PATENT
Docket No. 511582002700

CERTIFICATE OF FACSIMILE TRANSMISSION

I hereby certify that this correspondence is being facsimile transmitted to the United States Patent and Trademark Office on June 13, 2002.

Tami M. Procopio

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the application of:

Daniel E. H. AFAR, *et al.*

Serial No.: 09/389,000

Filing Date: 31 August 1999

For: PHELIX: A TESTIS-SPECIFIC PROTEIN EXPRESSED IN CANCER

Examiner: Minh-Tam B. Davis

Group Art Unit: 1642

24

DECLARATION OF PIA M. CHALLITA-EID

UNDER 37 C.F.R. § 1.132

Assistant Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

I, Pia M. Challita-Eid, declare as follows:

1. I have a Ph.D. in Microbiology from University of Southern California, did post doctoral work at University of California at Los Angeles, and was a faculty member at the University of Rochester. I have been practicing in the field of molecular biology for over 10 years. At Agensys, I am the Group Leader of Gene Discovery. In my position at Agensys, I have responsibility for evaluating the levels of expression of various genes in tissues. A copy of my *curriculum vitae* is enclosed as Exhibit A.

2. Our company, Agensys, is dedicated to discovery of proteins that are highly expressed in various tumor tissues as compared to normal tissues. The company approaches this discovery task by first identifying cDNAs which correspond to genes overexpressed in tumor

discovery task by first identifying cDNAs which correspond to genes overexpressed in tumor tissue using the technique of suppression subtractive hybridization (SSH). In this technique, cDNA from normal tissues is subtracted from cDNA from tumor tissues. Thereby, cDNA present in tumor tissues but not in normal tissues is isolated. Thus, on a gene-by-gene basis, this approach can indicate that a gene corresponding to the cDNA is overexpressed in tumor cells.

3. Typically, the next step is to utilize the sequence information obtained from SSH to obtain a full-length DNA clone which includes the entire open reading frame for the protein corresponding to this cDNA.

4. In addition, the level of expression of the corresponding gene is determined in various normal tissues and in various tumor tissues and tumor cell lines using the technique of Northern blotting, which detects production of messenger RNA. It is well known that the production of messenger RNA, that encodes the protein, is a necessary step in the production of the protein itself. Therefore, detection of high levels of messenger RNA by, for example, Northern blot, is a way of determining that the protein itself is produced.

5. Northern blotting is a detection method of relative levels of mRNA expression of a gene. It is procedure in which specific mRNA is measured using a nucleic acid hybridization technique. The signal is detected on an autoradiogram. The stronger the signal, the more abundant is the mRNA. For genes that produce mRNA that contains an open reading frame flanked by a good Kozak translation initiation site and a stop codon, in the majority of cases the synthesized mRNA codes for a protein. Kozak translation initiation sites are discussed in greater detail paragraph 7, below.

6. The evidence referred to in paragraphs 3, 4 and 5 above is consistent with the general knowledge in the art of molecular biology that, with rare exceptions, expression of a polynucleotide is predictive of expression of the corresponding protein. This is particularly true for mRNA with an open reading frame and a Kozak consensus sequence for translation initiation.

7. The consensus Kozak initiation site CCACC*ATG*G where the ATG start codon is italicized, refers to the "optimum" translation initiation sequence. A study by Peri and Pandey *Trends in Genetics* (2001) 17: 685-687, describes a study of over 1500 translation initiation sites in order to address the natural mRNA translation initiation. This study showed that the most authentic initiation sequence has 3 or more mismatches from the optimum consensus Kozak sequence CCACCATGG. The sequence of the translation initiation site of PHELIX, TCAACATGG, shows only 2 nucleic acid differences from the optimum Kozak consensus. Also, the translation initiation site of PHELIX contains a G at position +4, which has been shown to significantly augment translation efficiency (Kozak (1997) *Embo J* 16:2482-92). Altogether, these data demonstrate that the translation initiation site of PHELIX is functional and can initiate protein translation.

8. The Northern blot technique is used as a routine procedure (as compared to Western blotting, immunoblotting or immunohistochemistry) because it does not require the time delays involved in isolating or synthesizing the protein, preparing an immunological composition of the protein, eliciting a humoral immune response, harvesting the antibodies, and verifying the specificity thereof. All of these things can be done, but they take time, and the presence of mRNA on Northern blots, especially in comparative tissues, is a recognized indication that the protein itself will be produced.

9. I am familiar with the general practice of Northern blotting and interpretation, described above, being carried out, not only at Agensys, but also at other companies that seek to evaluate gene expression in various tumor and other tissues. The use of Northern blots as a means for evaluating protein production is universally accepted as reliable and is therefore widely practiced.

10. It is understood that the absolute levels of messenger RNA present and the amounts of protein produced do not always provide a 1:1 correlation. However, in those instances where the Northern blot has shown mRNA to be present, it is almost always possible, in my experience, when the time is taken to do so, to detect the presence of the corresponding

protein in the tissue which provided a positive result in the Northern blot. The levels of the protein compared to the levels of the mRNA may be disjunctive, but is would be inaccurate to say that there is no correlation between protein levels and mRNA levels as a general matter. In general, cells that exhibit detectable mRNA also exhibit detectable corresponding protein and *vice versa*. This is particularly true where the mRNA has an open reading frame and a good Kozak sequence.

11. Ironically, studies seeking to determine the overall pattern of correlation between mRNA and corresponding protein have started with displaying the protein fingerprint of a particular cell or tissue. For instance, an article by Anderson, L. and Seilhamer, J., *Electrophoresis* (1997) 18:533-537 (Exhibit B) describes such a study on a patient liver. A 2D gel was obtained to determine the pattern of proteins in the liver, and a cDNA library was used to determine the pattern for mRNA. The authors found that of 23 selected proteins which could be identified from the gel, mRNA for 19 were detected in the transcript images. Thus, in the vast majority of cases, there was both mRNA and protein present. The authors found that the levels of RNA units to protein units had a correlation coefficient of 0.48. As they state, this number is intriguingly close to the middle position between a perfect correlation (1.0) and no correlation whatever (0.0). Only a correlation coefficient of (0.0) would support a proposition that mRNA presence provides no indication of protein presence. The conclusion has to be that in the vast majority of instances, i.e., any correlation coefficient other than (0.0), where mRNA is present protein is also present. It is inaccurate to say that there is no correlation between mRNA expression and protein expression.

12. An article by Oh, J.M.C., *et al., Proteomics* (2001) 1:1303-1319 reports a database of protein expression in lung cancer. Again, the study sought to determine the correlation between mRNA and corresponding protein beginning with protein fingerprint display of a particular cell or tissue. Protein expression was evaluated using 2D gels and mRNA expression was evaluated using microarrays. The approach is suggested as a tool for evaluating, generically, the correlation between mRNA expression and protein expression. Clearly it is expected that the correlation will not be zero or the tool would not even be proposed.

13. I am aware that the Examiner has cited a publication by Fu, L., *et al., Embo. Journal* (1996) 15:4392 - 4401 which reports an extremely rare occurrence where there does appear to be zero production of any protein even in the presence of mRNA. This is for the specific protein p53. I am not familiar with any other instances where this occurs. This is an exception to the rule that there is at least some correlation between mRNA presence and protein production. This is supported by the publication itself; were this not an unusual occurrence, this lack of correlation would not merit publication at all.

14. In many cases, a reported lack of protein expression is due to technical limitations of the protein detection assay. For instance, the available antibody may only detect denatured protein but not native protein present in a cell. In other instances, the half-life of the protein is very short, thereby the steady-state protein levels are below detectable range. Short-lived proteins are still functional, and some have been previously described to induce tumor formation as shown in the article by Reinstein *et al. Oncogene 19: 5944-50.* In such situations, when more sensitive detection techniques are performed and/or other antibodies are generated, protein expression is detected. When studies fail to take these principles into account, they are likely to report artifactually lowered correlations of mRNA to protein.

15. A previous declaration has been submitted in this case to demonstrate that, at least in 293 cells, it is possible to produce the protein encoded by the PHELIX gene. As described in Dr. Hubert's declaration, this has been verified by producing antibodies raised against a 15-mer peptide designed from the PHELIX coding region. This demonstrates that in 293 cells, there is no translational inhibition to the production of protein.

16. The production of protein in the 293 cells shows conclusively that for those tumor cells, and by analogy for tumor cell lines where mRNA is also shown to be present, the PHELIX protein is present as well. The reason I conclude this is that in this experiment, when PHELIX mRNA was made, PHELIX protein was also produced and detected. This shows that the PHELIX mRNA is stable, functional and codes for a protein. And that the

translation initiation and termination sites of PHELIX are functional sites and lead to the production of a detectable PHELIX protein.

17. Most genes, when they produce mRNA that contains an open-reading frame flanked by a good Kozak translation initiation site and a stop codon, the synthesized mRNA code for a protein. Analysis of PHELIX shows a strong mRNA signal on Northern blot in cancer tissues, and the mRNA sequence shows an open-reading frame containing a good Kozak initiation site and a stop codon. Therefore, production of PHELIX protein is reasonably predicted based on this data.

18. In summary, the scientific community regards the presence of mRNA in cells is indicative of the production of protein. This is particularly true when the Northern data is strong and the mRNA has an open reading frame and a good Kozak sequence. It is understood that the correlation of mRNA and protein levels is not perfect, however, instances such as those in Fu, where protein is absent although mRNA is present at high levels, are a rare exception.

19. The use of positive Northern blots as indicative of and predictive of protein production is a recognized conclusion of scientists in this field.

20. I declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further, that these statements are made with the knowledge that willful, false statements and the like so made are punishable by fine or imprisonment or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Executed at Santa Monica CA on May 10, 2002.

Pia M. Challita-Eid

Curriculum Vitae

PIA M. CHALLITA-EID, PH.D

Personal information

Work address:	Agensys, Inc. 1545 17th Street Santa Monica, CA 90404
Email:	*pchallita@agensys.com*
Home address:	15745 Morrison Street Encino, CA 91436

Appointments:

Group Leader, Research Scientist III Gene Discovery	Agensys, Inc. October 2001-Present
Research Scientist II	Agensys, Inc. August 2000-Present
Assistant Professor in Medicine, Microbiology & Immunology	University of Rochester Cancer Center Hematology/Oncology Unit July 1998- June 2000
Senior Instructor	University of Rochester Cancer Center Department of Oncology January 1996- June 1998

Education:

B.S. Biology	American University of Beirut-Lebanon 1984-1987
M.S. Microbiology	American University of Beirut-Lebanon 1987-1989

Ph.D. Microbiology	University of Southern California Department of Microbiology January 1990 - June 1994
Advisor:	Donald B. Kohn, M.D., Associate, Professor Departments of Pediatrics and Microbiology Division of Research Immunology and Bone Marrow Transplantation Childrens Hospital of Los Angeles University of Southern California, California USA
Postdoctoral fellowship	University of California Los Angeles Department of Hematology-Oncology September 1994 - December 1995
Advisor:	Joseph D. Rosenblatt, M.D., Assistant Professor School of Medicine Department of Hematology-Oncology University of California, Los Angeles, California

Students and Research Associates Mentored:

Currently leading the Gene Discovery group of 6 research associates. Previous students and research associates mentored are listed below.

1. Skelton Diane, Research Associate, 1992-1994.
2. El-Khoueiry Anthony, Undergraduate student, Summer 1992 and 1993. Currently Fellow at the USC Medical Center.
3. Poles Tina, Research Associate, 1996-1998.
4. Mosammaparast Nima, Undergraduate student, June 1996 - September 1997. Currently enrolled in Medical School.
5. Zoric Bojan, Undergraduate student, June 1997-June 1998. Currently enrolled in Medical School.
6. Rimel BJ, Research Associate, June 1998-June 1999.
7. Vicki Houseknecht, Research Associate, June 1999 – June 2000.
8. Facciponte John, Graduate student in the Microbiology and Immunology Department at the University of Rochester, January 1998 - June 2000. Currently a graduate student at Roswell Park Cancer Center, Buffalo, NY.
9. Kyung Yi, Graduate Student in Microbiology, January 1999 - June 2000.
10. Anagha Joshi, Post-doctoral fellow, October 1999 – June 2000.

Patents:

In the last year, I have been involved in the filing of greater than 40 applications.

1) "Retroviral Vectors for Expression in Embryonic Cells", US5707865, issued date Jan. 13, 1998.
2) "Chimeric Proteins for the Stimulation of a Tumor-Specific Immune Response", application in progress.

Invited Presentations:

October 1994 "Retroviral Vector Expression in Murine Stem Cells". Department of Hematology-Oncology, UCLA Gene Therapy Program, Los Angeles, California.

October 1997 "Antibody Fusion Proteins for the Specific Recruitment and Activation of an Anti-Tumor immune Response". Childrens Hospital of Los Angeles, Los Angeles, California.

February 1998 Regional Cancer Center Consortium for Biological Therapy. Roswell Park Cancer Institute, Buffalo, New York.

July 1998 American Cyanamid Company. Lederle-Praxis Biologicals Division, Rochester, New York.

October 1999 "Monoclonal Antibody Technology in the Era of Genetic Engineering" Brazilian Meeting on Biosafety and Transgenic Products, Rio De Janeiro, Brazil.

June 1999 "Breast Cancer Research in the Era of Genetic Engineering", Breast Cancer Coalition of Rochester, Rochester, NY.

Awards:

Graduate Student Research Forum Award. Silencing of retroviral vectors after transduction of hematopoietic stem cells is associated with methylation. Graduate Student Research Forum Poster Session. USC Medical School, Los Angeles, California, 1993.

Presidential Award. Society of Biological Therapy, Pasadena, California, October 1997.

Merit Award. American Society of Clinical Oncology, California, May 1998.

Grants/Funds:

1) Jonsson Cancer Center Foundation/UCLA
Fellowship Seed Grant
Title: "Antigen Processing in Human Neural Crest Tumors"

Effective Dates: 11/1/95-10/31/96
Amount: $27,707

2) Rochester Area Foundation
Lucille B. Kesel Fund for the Advancenent of Cancer Research
Title: "Antibody Fusion Proteins for Eradication of Minimal Residual Disease"
Effective Dates: 1/1/98-12/31/98
Amount: $8,000

3) University of Rochester Cancer Center
Interim and Pilot Project Funding
P.I.: Joseph D. Rosenblatt, M.D.
Co-P.I.: Pia M. Challita-Eid, Ph.D.
Title: "Antibody Fusion Proteins for the Therapy of Cancer".
Effective Dates: 1/1/98-12/31/98
Amount: $25,000

4) Sinsheimer Scholar Award
Title: "Genetically-Engineered Chemokine Antibody Fusion Proteins for Breast and Ovarian Cancer Therapy"
Effective Dates: 7/1/98-6/30/01
Amount: $40,000/year

5) NIH/NCI
P.I.: Joseph D. Rosenblatt, M.D.
Co-P.I.: Pia M. Challita-Eid, Ph.D.
Title: "Recruitment and Activation of an Anti-tumor Response using Antibody-Fusion Proteins"
Effective Dates: 12/1/98-11/30/03
Amount: $191,046/year

6) NIH/NCI - Rapid Access to Intervention Development (RAID)
Title: "Preclinical Development of a B7.1 Anti-HER2/neu Antibody Fusion Protein"
Effective Date: Approved April, 1999
Amount: Not applicable

7) ACS Institutional grant
Title: "Chemokine Directed Targeting of Cytotoxic TALL-104 Cells"
Effective Dates: 9/1/99-8/30/00
Amount: $8,000

8) Breast Cancer Coalition of Rochester
Title: "Breast Cancer Research"
Date: 9/99
Amount: $1,000

Publications:

Gersuk GM, Westermark B, Mohabeer AJ, **Challita PM**, Pattamakom S, and Pattengale, PK. Inhibition of human natural killer cell activity by platelet-derived growth factor (PDGF). III. Membrane binding studies and differential biological effects of recombinant PDGF isoforms. *Scand J Immunol* 33: 521-532, 1991.

Gersuk GM, Carmel R, **Challita PM**, Rabinowitz AP, and Pattengale PK. Quantitative and functional studies of impaired natural killer (NK) cells in patients with myelofibrosis, essential thrombocytopenis, and polycythemia vera. I. A potential role for platelet-derived growth factor in defective NK cytotoxicity. *Nat Immun* 12: 136-151, 1993.

Challita PM, and Kohn DB. Lack of expression from a retroviral vector in murine hematopoietic stem cells is associated with methylation *in vivo*. *Proc Natl Acad Sci (USA)* 91: 2567-2571, 1994.

Krall W, **Challita PM**, Perlmutter L, Skelton D, and Kohn DB. Cells expressing human glucocerebrosidase from a retroviral vector repopulate macrophages and central nervous system microglia after murine bone marrow transplantation. *Blood* 83: 2737-2748, 1994.

Challita PM, Skelton D, Yu XJ, El-Khoueiry A, Yu X-J, Weinberg KI, and Kohn DB. Multiple modifications in *cis* elements of the long terminal repeat of retroviral vectors leads to increased expression and decreased DNA methylation in embryonic carcinoma cells. *J Virol* 69: 748, 1995.

Ucar K, Seeger RC, **Challita PM**,Watanabe CT, Yen TL, Morgan JP, Amado R, Chou E, McCallister T, Barber JR, Jolly DJ, Reynolds P, Gangavalli R, and Rosenblatt JD. Sustained cytokine production and immunophenotypic changes in human neuroblastoma cell lines transduced with a human gamma interferon vector. *Cancer Gene Therapy* 2: 171, 1995.

Lu Y, Planelles V, Palaniappan C, Li X, **Challita-Eid PM**, Amado R, Stephens D, Kohn DB, Bakker A, Day B, Bambara RA, and Rosenblatt JD. Inhibition of HIV-1 replication using a mutated tRNALys3 primer. *J Biol Chem* 272: 14523, 1997.

Challita-Eid PM, Penichet ML, Shin SU, Poles T, Mosammaparast N, Mahmood K, Slamon DJ, Morrison SL, and Rosenblatt JD. A B7.1-antibody fusion protein retains antibody specificity and ability to activate*via* the T cell costimulatory pathway. *J Immunol* 160: 3419-3426, 1998.

Challita-Eid PM, Abboud CN, Morrison SL, Penichet ML, Rosell KE, Poles T, Hilchey SP, Planelles V, and Rosenblatt JD. A RANTES- antibody fusion protein retains antigen specificity and chemokine function. *J Immunology* 161: 3729, 1998.

Challita-Eid PM, Rosenblatt JD, Day B, Rimel BJ and Planelles V. Inhibition of HIV-1 infection with a RANTES.IgG3 fusion protein. *AIDS Research and Human Retroviruses* 14:1617, 1998.

Mahmood K, Federoff HJ, **Challita-Eid PM**, Day B, Haltman M, Atkinson M, Planelles V, and Rosenblatt JD. Eradication of pre-established lymphoma using HSV amplicon vectors. *Blood* 93: 643, 1999

Penichet ML, **Challita PM**, Shin S-U, Sampogna S, Rosenblatt JD, and Morrison SL. In vivo properties of three human HER2/neu-expressing murine cell lines in immunocompetent mice. *Laboratory Animal Science* 49: 179-88, 1999.

Penichet ML, Dela Cruz JS, **Challita-Eid PM**, Rosenblatt JD, Morrison SL. A murine B cell lymphoma expressing human HER2/neu undergoes spontaneous tumor regression and elicits antitumor immunity. *Cancer Immunol Immunother* 49:649-62, 2001.

Hilchey SP, Rosebrough SF, Morrison SL, Rosenblatt JD, and **Challita-Eid PM**. Specific targeting and stimulation of in vivo anti-tumor response using a B7.1 T-cell costimulatory antibody fusion protein. *Manuscript in preparation.*

Select Abstracts and Presentations:

Challita PM, El-Khoueiry AB, and Kohn DB. Silencing of retroviral vectors after transduction of murine hematopoietic stem cells is associated with methylation. *Blood* 80 (10 Suppl. 1): 168a, 1992.

Challita PM, Cook C, Sender LS, and Kohn DB. Novel retroviral vectors for consistent expression after transduction into hematopoietic stem cells. Keystone Symposium on Gene Therapy, Keystone, Colorado, 1993.

Challita PM. Retroviral vector expression in murine stem cells. Presentation. Division of Hematology-Oncology, University of California Los Angeles, October, 1994.

Challita PM, Shin S-U, Penichet M, Mahmood K, Poles TM, Rosell KE, Abboud CN, Morrison SL, Rosenblatt JD. Novel Antibody Fusion Proteins for the Stimulation of a Tumor-Specific Immune Response. Keystone Symposium on Cellular Immunology and Immunotherapy of Cancer, Copper Mountain, Colorado, January 1997.

Penichet ML, **Challita PM**, Shin S-U, Slamon DJ, Rosenblatt JD, and Morrison SL. In vivo properties of two human her2/neu expressing murine cell lines in immunocompetent mice. Mutlidisciplinary Approaches to Cancer Immunotherapy, Bethesda, Maryland, June 1997.

Challita PM, Abboud CN, Rosell KE, Penichet ML, Poles T, Mahmood K, Morrison SL, and Rosenblatt JD. Characterization of a RANTES-antibody fusion protein for cancer immunotherapy. Mutlidisciplinary Approaches to Cancer Immunotherapy, Bethesda, Maryland, June 1997.

Horwitz S, Rosenblatt JD, Mosammaparast N, Poles T, Abboud CN, and **Challita PM**. Gene-modified EL4 cells expressing the chemokine RANTES protects from tumor growth and stimulates an anti-tumor cytotoxic T-lymphocyte response *in vivo*. Mutlidisciplinary Approaches to Cancer Immunotherapy, Bethesda, Maryland, June 1997.

Challita-Eid PM, Morrison SL, Penichet ML, Rosenblatt JD. Antibody-T cell costimulatory ligand fusion protein for the stimulation of a specific anti-tumor immune response. American Society of Hematology, San Diego, California, December 1997.

Challita-Eid PM, Abboud CN, Penichet ML, Rosell KE, Morrison SL, Rosenblatt JD. Antibody fusion proteins for the recruitment and activation of an anti-tumor immune response. American Association for Cancer Research. New Orleans, Louisiana, March 1998.

Challita-Eid PM, Hilchey Shannon P., and Rosenblatt Joseph D. An anti-HER2/neu RANTES fusion protein induces effector cell infiltration to the site of HER2/neu expressing tumors. AACR/NCI/EORTC Molecular Targets and Cancer Therapeutics, Washington DC, November 1999.

Facciponte JG, Rosenblatt JD, H.J.Federoff HJ, **Challita-Eid PM**. Herpes simplex virus (HSV) amplicon-mediated gene transfer of tumor associated antigens into bone marrow derived dendritic cells. Keystone Symposium on Cellular Immunity and Immunotherapy of Cancer, Santa Fe, New Mexico, January 2000.

A reassessment of the translation initiation codon in vertebrates

Suraj Peri and Akhilesh Pandey

More than two decades ago Marilyn Kozak proposed the scanning model of translation initiation, whereby translation is initiated at the first AUG codon that is in a particular context. In this article, we re-examine the context of initiator codons using a large dataset of curated human transcripts. We find that more than 40% of transcripts contain AUG codons upstream of the actual start codon and that most authentic AUGs contain three or more mismatches from the consensus sequence, CCACCaugG. Also, in a large fraction of transcripts, the sequences surrounding the initiator codon deviate more from the consensus than those surrounding upstream AUGs, indicating that translation initiation from downstream AUGs is more common than generally believed.

In this article, we re-examine the position requirement and the context dependence for an AUG codon to be used for translation initiation as proposed by Kozak[1]. The evidence in favor of the first-AUG rule is mainly derived from the fact that the first AUG was used in about 90–95% of cases in a study of several hundred vertebrate mRNAs (Ref. 2). The context refers to the nucleotide sequence surrounding the AUG (generally accepted as being CCACCaugG)[3,4]. This 'consensus' has been derived from observing the sequences surrounding the first AUG in exons as well as by mutagenesis studies. Although these mutagenesis studies provide data on what might constitute an 'optimal' sequence, they do not accurately predict whether a given AUG found in a natural mRNA transcript is likely to encode the initiator methionine.

As a result of the complete sequencing of the human genome, tens of thousands of novel genes have been identified, and their translation initiation sites need to be predicted[5,6]. Therefore, we decided to take another look at two of the major issues that underlie the scanning model of translation initiation. For our analysis, we have worked exclusively with well-characterized and annotated human mRNA sequences from the reviewed RefSeq dataset[7]. RefSeq is a project by the National Center for Biotechnology Information (NCBI) to create curated non redundant entries for each gene and contains predicted, provisional and reviewed entries for mRNAs and proteins. Reviewed entries are those where a human curator has performed an extensive manual verification, and by using these entries, we hope to avoid the annotation errors that otherwise abound in databases[8].

It has been proposed that certain classes of molecules, such as oncogenes, growth factors and receptors, are translated poorly and could contain a higher frequency of upstream AUG codons in their mRNAs as a mode of regulation[2]. We therefore subdivided our dataset into two classes: transcripts encoding cytosolic molecules, and those encoding proteins that are secreted or bound to the plasma membrane (i.e. those products with signal peptides). We first examined the nucleotide sequences surrounding their established initiator codons. Second, we inspected whether any AUGs exist upstream of the actual initiator codon in the 5′ untranslated regions (UTR) of these mRNAs. Last, because we found a significant number of transcripts with an upstream AUG, we examined whether there is any relationship between the size of the reported upstream 5′ UTR and the number of unused upstream AUGs.

Initiator codons in transcripts encoding cytosolic proteins

We studied the sequence context of AUGs from a dataset of 1534 reviewed transcripts encoding cytoplasmic proteins; the observed frequencies are shown in Table 1. When considered individually, the nucleotides that form the consensus CCACCaugG are found in 32–53% of transcripts. When only –3 and +4 positions are examined, only 46% of transcripts contain a purine (A or G) at –3 and a G at +4. Thus, over half of the transcripts differ from what are believed to be the most conserved nucleotide positions (–3 and +4) surrounding the AUG. We did not find that specific nucleotides occurred at position –5 with significantly higher frequency than random (P-value < 0.05). The degree of deviation of individual sequences from the consensus was also calculated and is discussed below.

Initiator codons in transcripts encoding cytokines, growth factors and receptors

The assignment of an AUG as an initiator methionine on the basis of genomic sequences can be quite contentious. Even when a protein sequence derived from a cloned cDNA is used, there can be disagreements on several issues. Therefore, we have taken a biological approach. Signal peptides are found at the amino termini of secreted factors such as cytokines and growth factors, as well as of type I transmembrane receptors that have their amino terminus located extracellularly[9,10]. These peptides are approximately 15–30 amino acids long and contain a stretch of hydrophobic amino acids. Several excellent programs are available that predict both the presence of signal peptide and the cleavage site[11,12]. Because signal peptides

Table 1. Frequency of nucleotides surrounding the initiator codon of transcripts encoding cytoplasmic proteins[a]

	-5	-4	-3	-2	-1	+1(A)	+2(T)	+3(G)	+4
A	17.07	21.77	47.00	30.37	20.99	100	0	0	19.36
T	19.29	10.16	5.60	10.56	6.12	0	100	0	13.36
G	31.42	28.61	37.15	18.57	27.64	0	0	100	53.38
C	32.20	39.43	10.23	40.48	45.24	0	0	0	13.88

[a]Sequences surrounding the initiator codon of 1534 manually reviewed RefSeq transcripts encoding cytoplasmic proteins. The frequency of occurrence of indicated nucleotides surrounding the initiator codon at positions –5 to +4 with respect to ATG is shown.

http://tig.trends.com 0168-9525/01/$ – see front matter PII: S0168-9525(01)02493-3

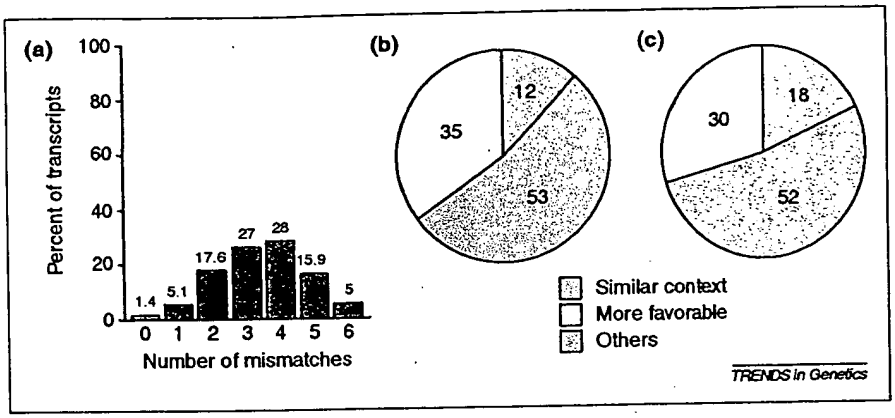


Fig. 1. Analysis of sequence contexts surrounding initiator codons and upstream unused AUGs. (a) Mismatch frequency of the nucleotides surrounding the initiator codon observed in natural transcripts as compared with the Kozak's consensus (CCACCaugG). This dataset is composed of 1534 transcripts encoding cytoplasmic proteins. The random probability of occurrence of no, one, two, three, four, five or six mismatches is 0.02%, 0.4%, 3.3%, 13.2%, 29.7%, 35.6% and 17.8%, respectively. (b) Comparison of the contexts of upstream unused AUGs with the authentic AUG, when CCACCaugG was considered as the optimal context (from a total of 2195 RefSeq transcripts). Yellow, the percentage of transcripts that contained at least one upstream AUG with fewer mismatches than the authentic AUG (i.e. the upstream AUG is in a more favorable context); green, the percentage of transcripts where all upstream AUGs were in a similar context (indicated by the same degree of mismatch from the consensus); orange, all other transcripts. (c) Comparison of the contexts of upstream unused AUGs with the authentic AUG when a purine (A or G) at −3 and a G at +4 was considered as the optimal context (from a total of 2195 RefSeq transcripts). The color code is the same as in (b).

can be recognized easily in these classes of proteins, assignment of the actual initiator methionine is obvious. Additionally, the cDNAs of most of these genes have been expressed in cells, validating the assignment of the signal peptides functionally.

We therefore compiled a list of 661 cytokines, growth factors and receptors, and then tabulated the nucleotide sequences surrounding the initiator methionine residue from transcripts encoding these proteins. As in the case above, only 41.8% of the transcripts contained a purine at −3 and a G at +4 (data not shown). The observations from this dataset essentially paralleled the results shown in Table 1, indicating that cytokines and growth factors do not contain any atypical motifs.

Frequency of initiator codons is not in agreement with the theoretical consensus 'CCACCaugG'

We next decided to investigate how often a real initiator methionine from our dataset is in agreement with the consensus and to express any deviation from the consensus as the number of mismatches observed. If the surrounding sequences were almost or entirely identical to Kozak's consensus, then one would expect to find most proteins with no or a single mismatch. However, if they were more randomly distributed, then the number of mismatches would be around three or four (because having exactly five or six mismatches is more constrained in terms of probability). Interestingly, we found that only ~24% of transcripts encoding cytosolic proteins had two or less mismatches compared with the consensus (Fig. 1a). This implies that a majority of transcripts contain initiator codons that are not in close agreement with Kozak's consensus sequence. The same phenomenon was seen when proteins with signal peptides were considered (data not shown).

Frequency of upstream AUGs

To determine how often the most 5′ AUG is used, we decided to inspect the transcripts for the presence of AUGs that were upstream of the initiator methionine. Here, we expected there would be no upstream AUGs in most of these cases. However, again to our surprise, we found that only slightly more than half of the transcripts contained no upstream AUG. In fact, 41% of transcripts had one or more, and 24% of genes had two or more upstream AUGs (data not shown). This means that, whatever the reason, the second, third or a further downstream AUG is chosen for translation initiation in these cases. Of course, if one were to assign the first AUG as the initiator methionine in these transcripts, the predicted open reading frame (ORF) and the length of the encoded protein would be erroneous. The lack of any significant difference in the distribution of cytosolic proteins and those with signal peptides indicates that the class of proteins coded for by the mRNAs is not a reliable indicator of atypical behavior of mRNAs.

It has been argued that it is the first AUG with a favorable context that is used for translation initiation. Therefore, we decided to compare the contexts of upstream unused AUGs with that of the authentic AUG in two ways. In the first method, we calculated the degree of mismatch for each of the upstream AUGs from the consensus, CCACCaugG, and compared it with the degree of mismatch of the authentic AUG. We divided the transcripts into three groups based on these results: (1) those that contained at least one upstream AUG in a more favorable context; (2) those where each of the upstream AUGs had a similar context; and (3) those where either all the upstream AUGs were in a less favorable context or some were in a less favorable context and others a similar context. In the second method, we divided the transcripts according to the degree of mismatch from a motif in which only two positions were considered as optimal; that is, a purine at −3 and a G at +4. The transcripts were again divided into three groups as in the first method. The object of this comparison was to identify the number of cases in which either a more favorable or similar AUG codon existed upstream of the authentic AUG. These two groups (categories 1 and 2, above) would therefore represent transcripts in which the first AUG that had a favorable context was not chosen for translation initiation.

The results according to the first method show that 35% of transcripts contained at least one upstream AUG that was in a more favorable context than the actual initiator, with 12% of transcripts containing upstream AUGs in a similar context to the authentic AUG (Fig. 1b). Figure 1c shows essentially similar results when only a purine at −3 and a guanine at +4 were considered as the optimal motif (according to the second method). Our analysis therefore reveals that in almost half the cases, there was at least one upstream unused AUG codon that was in

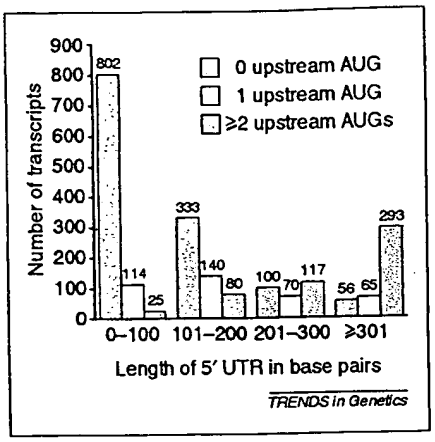


Fig. 2. The length of 5′ untranslated region (UTR) in basepairs plotted against the number of transcripts having no, one or ≥ two AUGs upstream of the actual translation start codon. 2195 RefSeq transcripts were analyzed. No significant difference was found between datasets of transcripts encoding cytoplasmic proteins and those with signal peptides.

a similar or better context than the authentic AUG.

The length of the 5′ untranslated region is related to the number of upstream unused AUGs

While we were performing this analysis, we were intrigued by the fact that if the 5′ UTR was long, the transcripts invariably contained upstream AUGs that were not used. We therefore decided to investigate this systematically. Figure 2 shows a histogram of the number of transcripts with no, one and ≥ two upstream AUGs plotted against the length of the 5′ UTR. Most of the transcripts (85%) with 100 bp or less of 5′ UTR sequence do not contain any upstream AUGs, and only 2.6% contain two or more upstream AUGs. Quite the opposite is observed for transcripts with 5′ UTRs longer than 300 bp. In this case, 70% of the transcripts contain two or more upstream AUGs, and only 13.5% contained no upstream AUGs. As the length of the 5′ UTR increases, the number of transcripts with no upstream AUGs decreases and the number of transcripts with unused upstream AUGs increases. We see the same phenomenon in both classes of our dataset, indicating that it is not dependent on the type of protein being studied. Considering that the average length of the 5′ UTR for genes in the human genome is 300 bp (Ref. 6), our data suggests that one has to be quite careful with the 'first AUG' rule because it is probable that the first AUG is not being used in a significant number of transcripts.

Conclusions

Our analysis essentially focused on testing whether there is any consensus around the initiator codon in transcripts encoding known proteins. Transcripts that encode well-studied proteins provide a more applicable dataset, as these proteins are not predictions and have been worked on by scores of investigators worldwide. They are also good candidates to test the predictive value of Kozak's criteria when considering assignment of a given AUG in the transcript of a newly discovered gene. Our analysis shows that a large number of transcripts contain AUGs upstream of the actual translation start site, many of which are in a more favorable context than the codon used for translation initiation. Furthermore, our data shows that most of the AUGs used for translation initiation deviate significantly from Kozak's consensus. It is possible that mechanisms such as leaky scanning, reinitiation or internal initiation of translation have a much greater role than previously imagined[13–16]. In support of this idea, a growing number of transcripts have recently been reported to undergo internal initiation[17–19]. For the purposes of gene prediction and identification of translation start sites from genomic DNA or cDNA sequences, it is better to use homology-based alignments across protein families or across species to identify the initiator codon correctly, instead of relying solely on the most upstream AUG and its context.

Acknowledgements

Work at the Center for Experimental Bioinformatics is supported by a generous grant from the Danish National Research Foundation. A.P. is supported by a Howard Temin Award (CA75447) from the NCI, National Institutes of Health. A.P. thanks Harvey Lodish and Matthias Mann for their support. S.P. has a grant from the Plasmid Foundation, Denmark, and acknowledges M. Srihari (M.S. University of Baroda) for helpful discussions and B.B. Chattoo for his encouragement.

References

1 Kozak, M. (1978) How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* 15, 1109–1123
2 Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* 15, 8125–8148
3 Kozak, M. (1984) Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12, 857–872
4 Kozak, M. (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* 9, 5233–5262
5 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
6 International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
7 Pruitt, K.D. *et al.* (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 16, 44–47
8 Peri, S. *et al.* (2001) Common pitfalls in bioinformatics-based analyses: look before you leap. *Trends Genet.* 17, 541–545
9 Spiess, M. (1995) Heads or tails – what determines the orientation of proteins in the membrane. *FEBS Lett.* 369, 76–79
10 Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein. Chem.* 54, 277–344
11 Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *ISMB* 6, 122–130
12 Claros, M.G. *et al.* (1997) Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.* 7, 394–398
13 Slusher, L.B. *et al.* (1991) mRNA leader length and initiation codon context determine alternative AUG selection for the yeast gene *MOD5*. *Proc. Natl. Acad. Sci. U. S. A.* 88, 9789–9793
14 Jackson, R.J. and Kaminski, A. (1995) Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *RNA* 1, 985–1000
15 Liu, C.C. *et al.* (1984) Initiation of translation at internal AUG codons in mammalian cells. *Nature* 309, 82–85
16 Gray, N.K. and Wickens, M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.* 14, 399–458
17 Vagner, S. *et al.* (1995) Alternative translation of human fibroblast growth factor 2 mRNA occurs by internal entry of ribosomes. *Mol. Cell. Biol.* 15, 35–44
18 Sehgal, A. *et al.* (2000) The chicken c-Jun 5′ untranslated region directs translation by internal initiation. *Oncogene* 19, 2836–2845
19 Coldwell, M.J. *et al.* (2000) Initiation of Apaf-1 translation by internal ribosome entry. *Oncogene* 19, 899–905

Suraj Peri
Protein Interaction Laboratory,
Center for Experimental Bioinformatics,
University of Southern Denmark,
Odense M, Denmark.

Akhilesh Pandey*
Whitehead Institute for Biomedical Research,
Cambridge, MA 02142, USA.
*e-mail: pandey@bmb.sdu.dk

The EMBO Journal Vol. 16 No. 9 pp. 2482-2492, 1997

Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6

Marilyn Kozak

Department of Biochemistry, University of Medicine and Dentistry of New Jersey, 675 Hoes Lane, Piscataway, NJ 08854, USA

Abstract

A primer extension (toeprinting) assay was used to monitor selection by ribosomes of the first versus the second AUG codon as a function of introducing mutations on the 3′ side (positions +4, +5 and +6) of the first AUG codon. Six different flanking codons starting with G (GCG, GCU, GCC, GCA, GAU and GGA) strongly augmented selection of AUG#1 when compared with matched mRNAs that had A or C instead of G in position +4. Augmentation by G in position +4 failed only when it was combined with U in position +5, as in the sequence augGUA. In contrast with the usual enhancing effect of introducing G in position +4, most mutations in position +5 had no discernible effect, as shown with the series augANA (where N = C, A, G or U) and the series augCNA. AUG codon recognition was also unaffected by mutations in position +6, as shown by testing four mRNAs that had augCCN as the start site. Thus the primary sequence context that augments the recognition of AUG start codons does not appear generally to extend beyond G in position +4. When the toeprinting assay was used with mRNAs that initiate translation at CUG instead of AUG, cugGAU was not recognized better than cugGGU, contradicting the hypothesis that initiation at non-AUG codons might be favored by A instead of G in position +5.

Keywords: initiation codon context/mRNA structure/protein synthesis/scanning model/translation

Introduction

Eukaryotic ribosomes appear to select the start site for translation by a scanning mechanism. The working hypothesis is

that the small, 40S ribosomal subunit, carrying Met-tRNA$_i^{met}$ and various initiation factors, engages the mRNA at the capped 5′ end and migrates linearly until it encounters the first AUG codon. At the AUG codon, which is recognized by base pairing with the anticodon in Met-tRNA$_i^{met}$, the 40S ribosomal subunit stops, the 60S subunit joins and the 80S ribosome is poised to start protein synthesis. Evidence for this scanning mechanism and for the corollary first-AUG rule is summarized elsewhere (Kozak, 1989a, 1992, 1995).

In higher eukaryotes, sequences flanking the AUG codon modulate its ability to halt the scanning 40S ribosomal subunit. One of the modulating elements is the GCCACC motif in positions -6 to -1, immediately preceding the AUG codon (Kozak, 1987). Mutations that weaken adherence to this consensus motif, especially mutations that substitute a pyrimidine for the A in position -3, cause some 40S ribosomal subunits to bypass the first AUG and to initiate instead at the next AUG downstream (Kozak, 1986a, 1989b; Lin *et al.*, 1993; Ossipow *et al.*, 1993). This context-dependent 'leaky scanning' has also been seen when the highly conserved G in position +4, immediately following the AUG codon, is mutated (Kozak, 1986a, 1989b). Deviations in one or both of these key positions, and the resulting leaky scanning, seem to account for the ability of certain mRNAs to produce two proteins by initiating translation from the first and second AUG codons (Kozak, 1986b, 1991).

Two recent studies have raised the possibility that context effects on initiation might extend into the coding domain beyond position +4. In one case, initiation at GUG appeared to be more efficient when the second codon was GAU instead of GUA (Boeck and Kolakofsky, 1994). A companion study by Grünert and Jackson (1994) reported similarly that initiation at an AUG or CUG start codon was favored by A in position +5 and U in position +6.

However, documenting the involvement of these or other nucleotides on the 3′ side of the initiator codon might be complicated by the fact that mutations introduced in these positions of the mRNA may change the amino acid sequence of the encoded polypeptide. This could be a problem because the identity of the amino acid adjacent to the N-terminal methionine can affect post-translational modifications which, in turn, can affect protein turnover. To circumvent possible complications from post-translational events, an assay that directly monitors ribosome-mRNA initiation complexes was used in the present study to reinvestigate the question of whether nucleotides in positions +4, +5 and +6 affect the recognition of initiator codons.

Correct definition of the context requirements for initiation is important for predicting translational start sites, which is an important aspect of interpreting cDNA sequences (Kozak, 1996).

Results

Preliminary test of mutations in positions +4, +5 and +6

The mRNAs used for these experiments have two start codons and two open reading frames (ORFs), as outlined in Figure 1. ORF1, which extends from AUG#1 to a UAA codon overlapping Leu45 in the chloramphenicol acetyltransferase (CAT) coding sequence, encodes a 70 amino acid polypeptide with a molecular mass of 8 kDa. This polypeptide is designated p8^{out} (meaning out-of-frame with respect to CAT) or simply p8. ORF2 initiates with AUG#2, which is in-frame with the downstream CAT coding sequence. ORF2 thus encodes a 240 amino acid polypeptide (the 219 amino acid CAT protein with a 21 amino acid N-terminal extension), with a molecular mass of 28 kDa. The product of ORF2 is designated p28^{precat} or simply p28. Because the sequence preceding AUG#1 includes U in position -3, which is suboptimal, leaky scanning should allow these mRNAs to produce both polypeptides: p8 from AUG#1 and p28 from AUG#2. The control mRNA in Figure 2A (lane 9) illustrates how this leaky scanning can be modulated by changes in

context. Because AUG#1 in the control has the optimal A in position -3, this mRNA produces a much higher yield of p8, and a much lower yield of p28, than any other mRNA in this series. This fits with previous studies of mutations involving sequences on the 5′ side of the AUG codon (Kozak, 1986a, 1989b).

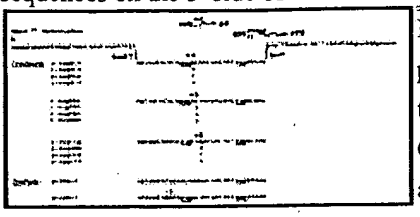


Fig. 1. Sequences of mRNAs used to study the effects of varying the nucleotide in positions +4, +5 and +6. The sequence in the top line is common to all mRNAs in this series. The 3′ end of this sequence indicated by an ellipsis (. . .) leads to the CAT coding domain (Kozak, 1989b). Not shown is the m7G cap at the 5′ end of all mRNAs. Mutations in the indicated positions of particular mRNAs were introduced around the first AUG codon, which initiates translation of an 8 kDa polypeptide (p8). In a different, overlapping reading frame, the second AUG codon initiates translation of a 28 kDa polypeptide (p28) which is an N-terminally extended version of CAT. Because of the suboptimal context preceding AUG#1 (notably the presence of U rather than A in position -3), some ribosomes would be expected to reach AUG#2 by leaky scanning. Thus, each of the 12 test constructs should direct translation of both p8 and p28: The improved context (A in position -3) in the control p-Con-1 should strongly shift translation in favor of p8. With the other control, p-Con-0, p28 should be the sole translation product because the upstream AUG codon is absent. Notice that mRNAs are named by stating the three bases following the first AUG codon.
[View Larger Version of this Image (13K GIF file)]

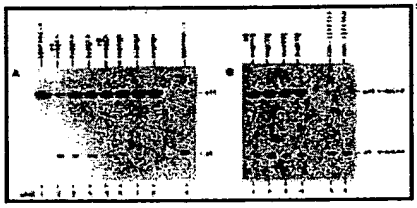


Fig. 2. Translation of mRNAs that vary in positions +4, +5 and +6 flanking the first AUG codon. The autoradiograms show [^{3}H]leucine-labeled proteins produced in a rabbit reticulocyte translation system using mRNAs that have point mutations in positions +4 and +5 (**A**) or position +6 (**B**). The mutations identified above each lane were introduced around AUG#1, which initiates translation of p8. p28 results from initiation at the invariant AUG#2. Figure 1 gives the 5′ end sequences of these mRNAs in full. A control mRNA that lacks AUG#1 (p-Con-0 in lane 1 of A) produced only p28. For the control mRNA p-Con-1 (lane 9 in A; lane 5 in B), the context around AUG#1 was improved by changing position -3 from U to A, thus enhancing synthesis of p8 and greatly reducing synthesis of p28. In (B), the slight residual translation of p28 evident in lane 5 was abolished in lane 6 by introducing downstream the structure-prone sequence 8336, which is thought to slow scanning and thus augment recognition of AUG#1 (Kozak, 1990a). This is shown only as an illustration, inasmuch as all the other mRNAs used in this figure contained the unstructured sequence 8335 at the *Bam*HI site. The conditions used for translation (protein accumulation assay) and subsequent fractionation by polyacrylamide gel electrophoresis are described in Materials and methods.
[View Larger Version of this Image (39K GIF file)]

The present study asks whether mutations on the 3′ side of AUG#1 can also modulate the selection of translational start sites. As shown in Figure 2A (lanes 2-8), the yield of p8 initiated from AUG#1 indeed varied at least 5-fold when point mutations were introduced in positions +4 or +5. However, the scanning mechanism predicts that if, for example, p-augAAA really supports initiation better than p-augAUA, as suggested by the 5-fold higher yield of p8 in lane 6 versus lane 8, then the yield of p28 should be proportionately lower in lane 6. That prediction is not met. Instead, the only mRNA in the test series that shows both elevated p8 synthesis and reduced p28 synthesis is p-augGCA (lane 2), the construct that has G instead of U, C or A in position +4.

With the other mRNAs tested in Figure 2A, a possible explanation for the variable yield of p8 without concomitant reduction of p28 is that mutations in positions +4 and +5, which change the subterminal amino acid, thereby alter the turnover of polypeptide p8. In this case, the amount of radiolabeled p8 that accumulates during the hour-long incubation would not reflect the efficiency of initiation at AUG#1 accurately. To circumvent this potential problem, the mRNAs used

in Figure 2A were retested using a direct initiation assay.

To examine the effects of mutations in position +6, I chose a codon that specifies the same amino acid regardless of which base occurs in position +6. Thus the N-terminal sequence of the nascent polypeptide is Met-Pro when translation initiates at augCCG, augCCU, augCCC or augCCA. Among these four mRNAs there was no significant difference in the yield of p8 in a standard translation assay (Figure 2B, lanes 1-4). These mRNAs were also retested using the initiation assay described next.

Direct analysis of AUG codon recognition using mRNAs with mutations in positions +4, +5 and +6

By using a reticulocyte lysate supplemented with sparsomycin and cycloheximide to inhibit elongation (see Materials and methods), initiation complexes accumulate in which the ribosome is held at the AUG codon. The particular AUG start site can be identified by using a primer extension inhibition assay in which a ^{32}P-labeled deoxyoligonucleotide primer, annealed to the mRNA downstream from all potential initiator codons, is extended by reverse transcriptase up to the 3′ edge of the bound ribosome. Figure 3 outlines how the assay works in principle.

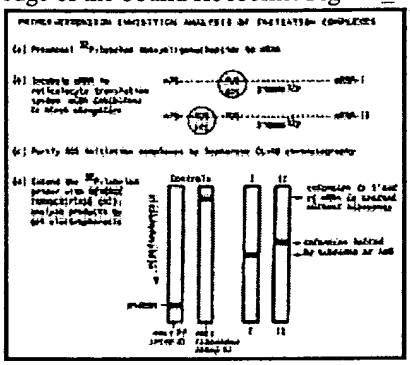


Fig. 3. Schematic representation of a primer extension assay for mapping the position of ribosomes on mRNA. The unextended ^{32}P-labeled primer, represented by the wide black line in step (b), is shown near the bottom of the polyacrylamide gel in step (d). Extension of the primer with reverse transcriptase in the absence of bound ribosomes proceeds to the 5′ end of the mRNA. If ribosomes are allowed to bind to the mRNA before the addition of reverse transcriptase, primer extension halts prematurely; the exact size of the extension product(s) reveals which AUG codon(s) were selected, taking into account that the leading edge of an 80S ribosome extends ~15 nucleotides 3′ of the AUG codon (Kozak and Shatkin, 1977). The basic design of this 'toeprinting' assay was developed by Hartz *et al.* (1988) for studies with prokaryotic ribosomes. The primer used in the present studies was 23 nucleotides long, the full-length extension product was 184 nucleotides and the extension inhibition products obtained when a ribosome was bound at AUG#1 or AUG#2 were 123 and 109 nucleotides, respectively.
[View Larger Version of this Image (35K GIF file)]

Two control reactions in Figure 4 illustrate how the assay works in practice. With the control mRNA p-Con-1 in which AUG#1 resides in a nearly optimal context (ACAaugG, see Figure 1), one prominent primer extension product is evident in Figure 4A (lanes 11 and 12) and the size of this product indicates that it derives from ribosomes bound at AUG#1, the start codon for p8. This primer extension product was absent when p-Con-0 mRNA was used for ribosome binding (Figure 4A, lanes 1 and 2), consistent with the fact that p-Con-0 lacks the upstream AUG codon (see Figure 1). With p-Con-0 the toeprinting assay maps ribosomes instead at the p28 start codon. (The p28 start site is labeled AUG#2 in Figure 4 because it is the second start codon in all mRNAs except p-Con-0.)

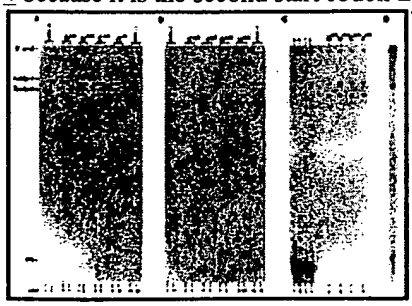


Fig. 4. Primer extension analysis of ribosome-mRNA complexes. Initiation at the first and second AUG codons was monitored as a function of introducing mutations around AUG#1. The assay is explained in Figure 3. The primer (PR) and primer extension products are labeled along the left margin. (**A** and **B**) The mRNAs used in lanes 3-10 varied only in position +4 or position +5, as indicated at the top of each panel. The sequences of these mRNAs as well as the two control transcripts (lanes 1, 2, 11 and 12) are depicted in full in Figure 1. Adjacent bracketed lanes show that, with a given mRNA, the ratio of initiation at AUG#1 versus AUG#2 was the same when low (lanes 1, 3, 5, 7, 9 and 11) and 3-fold higher (lanes 2, 4, 6, 8, 10 and 12) levels of initiation

complexes were analyzed. Because of small variations in the amount of radioactivity applied to each lane, the important comparison is not the intensity of the AUG#1 band from lane to lane, but the ratio of AUG#1 to AUG#2 in each lane. These ratios are given in Table I. (C) Toeprint analyses with mRNAs that differed in position +6, as indicated above lanes 5-8. Lanes 1-4 display the complementary strand sequence of p-augCCA mRNA. A series of black dots within these sequencing lanes highlight the C-A-T bands that correspond to the first, second and (silent) third AUG codons. When the electrophoresis was run for longer, the foreshortened primer extension products caused by bound ribosomes could be mapped, by reference to the sequencing lanes, 15-16 nucleotides downstream from the first and second AUG codons. In the absence of ribosomes, the primer was extended all the way to the 5′ end of the mRNA, as shown in a control reaction (**D**).

[View Larger Version of this Image (78K GIF file)]

The rest of Figure 4A tests the effects of introducing mutations in position +4 flanking AUG#1. Because all four test transcripts (the first four mRNAs in Figure 1) have U instead of the optimal A in position -3, some ribosomes are able to reach AUG#2 by leaky scanning. The question is whether the ratio of initiation at AUG#1 versus AUG#2 differs among these four mRNAs which are identical except for position +4. Quantitation of the data from Figure 4A (Table I, measurement 1, entries 1-4) indeed shows a 2.6-fold shift in favor of AUG#1 when that codon is followed by G in position +4 (henceforth written G^{+4}). There was no real hierarchy among the other three nucleotides in position +4.

Table I. Effects of mutations in positions +4, +5 and +6 as monitored by primer extension analysis of ribosome-mRNA initiation complexes

[View Table]

The toeprinting experiment was repeated in Figure 4B using four mRNAs that were identical except for position +5. Quantitation of these results showed no significant shift in the AUG#1/AUG#2 ratio (Table I, measurement 1, entries 5-8). Thus there was no evidence that the nucleotide in position +5 affects the selection of translational start sites. Nor was there any significant effect when mutations in position +6 were tested (Figure 4C; Table I, measurement 1, entries 9-12).

Because leaky scanning in cell-free translation systems was shown previously to be sensitive to the concentration of Mg^{2+} (Kozak, 1989b, 1990b), I repeated the test of mutations in positions +4, +5 and +6 at three different Mg^{2+} concentrations. The results of these toeprinting assays are shown in Figure 5 and the quantitation is given in Table I (measurements 2, 3 and 4). As reported previously, when a given mRNA is tested at different Mg^{2+} concentrations, the tendency to scan past AUG#1 and initiate instead at AUG#2 increases as the Mg^{2+} concentration is decreased. This can be seen in Figure 5A, for example, by comparing the translation of p-augGCA in lanes 1, 5 and 9. The point is sustained by comparing any other mRNA in Figure 5A at low, medium and high concentrations of Mg^{2+} (e.g. p-augUCA in lanes 2, 6 and 10; p-augCCA in lanes 3, 7 and 11; or p-augACA in lanes 4, 8 and 12).

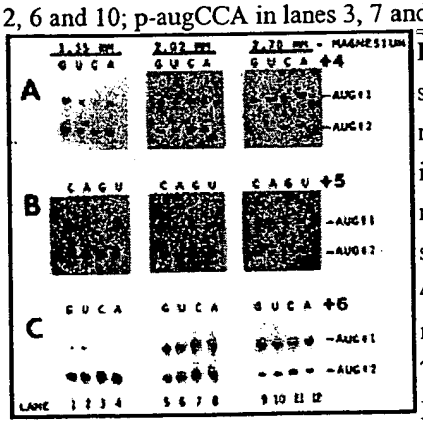


Fig. 5. Selection of AUG start sites as a function of magnesium concentration and sequence variation following AUG#1. (**A**) Twelve toeprinting reactions using mRNAs that differed from one another only in position +4 following AUG#1. The identity of the base in position +4 is marked above each lane. The primer extension reactions were carried out after 4 min incubation in a reticulocyte translation system in which the magnesium concentration had been adjusted to 1.35 (lanes 1-4), 2.02 (lanes 5-8) or 2.70 mM (lanes 9-12). The experiment was repeated using mRNAs that differed from one another only in position +5 (**B**) or position +6 (**C**). The mRNA sequences are given in full in Figure 1. Autoradiograms (similar to Figure 4) have been cropped. Quantitation of these results is given in Table I.

[View Larger Version of this Image (77K GIF file)]

The real purpose of this experiment was to compare, at a given concentration of Mg^{2+}, the translation of four mRNAs that differ in a single position downstream of AUG#1. In Figure 5A (lanes 1-4) this four-way comparison reveals that G is the only nucleotide in position +4 that augments recognition of AUG#1 at low Mg^{2+}, and that conclusion holds at moderate (lanes 5-8) and high (lanes 9-12) Mg^{2+} concentrations. In Figure 5B, a similar experiment using four mRNAs that differ only in position +5 shows that, at any given Mg^{2+} concentration, AUG#1 is recognized with equal efficiency irrespective of nucleotide changes in position +5. Figure 5C shows that, at any given Mg^{2+} concentration, recognition of AUG#1 is indifferent also to the nucleotide in position +6. The conclusion from this analysis is that G^{+4} appears to be the only nucleotide on the 3′ side of the AUG codon that augments initiation.

Distinguishing between particular codon effects and generalized context effects

Although G^{+4} augments AUG codon recognition under a variety of reaction conditions, as shown above, in all those studies the G in position +4 was part of the codon GCA. To determine if the augmentation is attributable specifically to G^{+4} or if it is the flanking codon GCA that happens to favor initiation, I tested mRNAs that had six different GNN codons adjacent to AUG#1. In the toeprint analyses shown in Figure 6A, each mRNA was compared with a matched construct that had C or A instead of G in position +4. Quantitation of the results (Table II) reveals that AUG#1 was indeed recognized ~3-fold better in five out of six cases where G^{+4} was the flanking nucleotide. Since five different flanking codons starting with G (GCG, GCU, GCC, GCA and GAU) strongly augmented the recognition of AUG#1, it seems reasonable to attribute the enhancement to the G residue in position +4 rather than to a particular flanking codon.

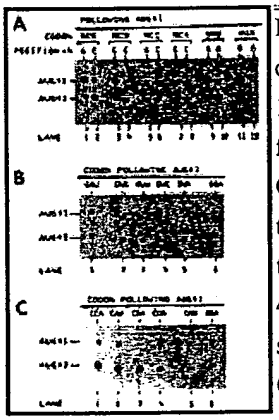


Fig. 6. Additional testing of various codons following AUG#1. Primer extension assays were carried out using a reticulocyte translation system with the Mg^{2+} concentration adjusted to 1.7 mM. The codon flanking AUG#1 is identified above each lane of the gel. Except for this first codon variation, the mRNA sequences were as given in Figure 1. **(A)** The positive effect of G^{+4} is seen with a variety of codons flanking AUG#1. The first eight lanes, compared two at a time, show the shift in initiation when AUG#1 is followed by G versus C in position +4. Thus the flanking codons are GCG versus CCG in lanes 1 and 2, GCU versus CCU in lanes 3 and 4, GCC versus CCC in lanes 5 and 6 and GCA versus CCA in lanes 7 and 8. The last four lanes show the shift when AUG#1 is followed by G versus A in position +4. The flanking codons are GAU, lane 9; AAU, lane 10; GUA, lane 11; and AUA, lane 12. Quantitation of these results is given in Table II. **(B)** The stimulatory effect of G^{+4}, evident when AUG#1 is followed by GAU (lane 1) or GGA (lane 6), fails when the flanking codon is GUN (lanes 2-5). **(C)** Variations in position +5 do not significantly affect recognition of AUG#1, as shown with CNA as the flanking codon (lanes 1-4). The mRNAs used in lanes 5 and 6 are controls. Warming during electrophoresis slightly retards the mobility of samples near the edges of the gel.

[View Larger Version of this Image (54K GIF file)]

Table II. The positive effect of G in position +4 occurs with a variety of codons flanking AUG#1

[View Table]

In Figure 6A, augGUA was the only mRNA in which G^{+4} unexpectedly failed to enhance initiation. To determine whether it is specifically the flanking codon GUA that disfavors initiation or whether the 3′ sequence GU somehow undermines recognition of the preceding AUG codon, I tested initiation at augGUG, augGUU and augGUC along with augGUA. Figure 6B (lanes 2-5) shows equally poor recognition of AUG#1 with all four constructs in this series. Thus the usual

stimulatory effect of G^{+4}, seen in Figure 6B with the control transcripts augGAU and augGGA (lanes 1 and 6), fails for some reason when G^{+4} is followed by U^{+5}.

Although U in position +5 prevents the usual stimulatory effect of G^{+4}, U in position +5 is not generally deleterious. Thus, augA**U**A was not recognized less efficiently than augA**C**A, aug A**A**A or augA**G**A in Figure 5B. The point is confirmed in Figure 6C where augC**U**A (lane 4) was recognized as efficiently as augC**C**A, augC**A**A or augC**G**A (lanes 1-3).

Effects of mutations flanking a CUG start codon

In view of some earlier reports about effects of downstream mutations (Boeck and Kolakofsky, 1994; Grünert and Jackson, 1994), it seemed useful to retest some of the foregoing conclusions with mRNAs that initiate translation at a non-AUG codon. In the mRNAs depicted in Figure 7A, CUG replaces AUG#1 as the start codon for p8. When these mRNAs were used as templates in a standard translation assay, some [^{3}H]leucine-labeled p8 was produced (Figure 7B, lanes 2-5), albeit less than with AUG as the start site for p8 (Figure 7B, lanes 1 and 6). That AUG as the p8 start codon is much stronger than CUG is also evident from the greater inhibition of p28 synthesis in lanes 1 and 6 compared with lanes 2-5 in Figure 7B. The complete absence of p8 when the CUG codon was mutated (Figure 7C, lane 1) confirms that CUG is the source of p8 in lanes 2-5.

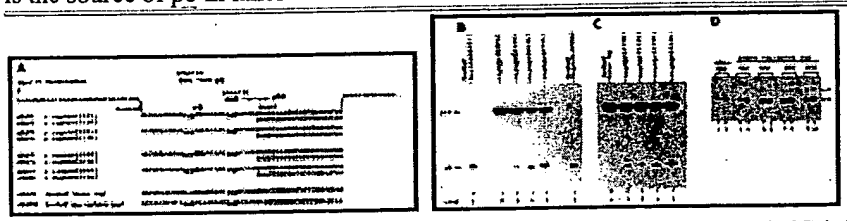


Fig. 7. Effects of mutations in positions +4 and +5 on recognition of a CUG initiator codon. (**A**) Sequences of the mRNAs used for translation. The major difference from Figure 1 is that AUG#1 has been replaced here by a CUG start codon. To compensate for the weakness of the CUG codon, the preceding sequence includes the optimal A in position -3. Each construct was tested with an unstructured sequence (8335) and a secondary structure-forming sequence (8336) downstream. (**B** and **C**) Autoradiograms of [^{3}H]leucine-labeled proteins resulting from translation under standard conditions (2 mM Mg^{2+}) of the mRNA indicated above each lane. The autoradiogram in (C) was exposed twice as long as that in (B). (**D**) Cropped autoradiograms of toeprinting reactions carried out with p-cugAGU mRNA (lanes 3 and 4), p-cugAAU (lanes 5 and 6), p-cugGGU (lanes 7 and 8) and p-cugGAU (lanes 9 and 10). Lanes 1 and 2 show a control reaction with mRNA #9 which lacks the upstream CUG start site. For each mRNA, adjacent bracketed lanes show toeprinting reactions carried out using the first two ^{32}P-containing fractions eluted from the Sepharose CL-4B column. Because of difficulty in synchronizing the collection when several columns are run at the same time, the first fraction (lanes 1, 3, 5, 7 and 9) contains less radioactivity in some cases. Consequently, the mRNAs are most easily compared by focusing on lanes 2, 4, 6, 8 and 10, where equal radioactivity was applied.
[View Larger Versions of these Images (22 + 36K GIF file)]

Because earlier studies showed that appropriately positioned downstream secondary structure aids the recognition of weak initiator codons (Kozak, 1990a), I tested the CUG mRNAs with both an unstructured sequence (oligonucleotide 8335) and a structure-forming sequence (oligonucleotide 8336) downstream. The inclusion of oligonucleotide 8336 significantly elevated initiation from the CUG codon in Figure 7B (compare the yield of p8 in lane 4 versus lane 2, or in lane 5 versus lane 3); therefore, this downstream sequence was retained in the mRNAs used in Figure 7C and D.

In Figure 7C, I examined the effects of mutations in position +5 flanking the CUG start codon for p8. I specifically tested A versus G in this position because Grünert and Jackson (1994) reported the biggest effect when A^{+5}, which they

considered optimal, was changed to G^{+5}. Although I too observed a higher yield of p8 with the mRNA that initiates at cugGAU instead of cugGGU (lanes 4 and 5 in Figure 7B and C), closer inspection of the results argues against concluding that A^{+5} enhances initiation at CUG. Notice, for example, that the yield of p8 was not augmented by A in position +5 when the flanking codons were AGU versus AAU (Figure 7C, lanes 2 and 3) instead of GGU and GAU.

Because differences in protein stability might distort the results of the protein accumulation assays in Figure 7B and C (for example, N-terminal acetylation might stabilize the form of p8 initiated from cugGAU), the critical test was whether mutations in position +5 would affect recognition of the CUG codon when initiation was monitored directly, using the primer extension assay. As shown in Figure 7D, although G in position +4 augmented recognition of the upstream CUG codon (compare cugAGU with cugGGU, lanes 4 and 8; or compare cugAAU with cugGAU, lanes 6 and 10), there was no convincing difference between matched mRNAs that had G versus A in position +5 (compare cugAGU with cugAAU in lanes 4 and 6; or compare cugGGU with cugGAU in lanes 8 and 10).

Notice that the extended context in these CUG-containing mRNAs (GACAUAcugRRU) is the same as that used by Grünert and Jackson (1994).

Discussion

The optimal context for initiation does not extend beyond G in position +4

These experiments refute the suggestion that the recognition of initiator codons is strongly favored by A in position +5 and U in position +6, as proposed by Boeck and Kolakofsky (1994) and Grünert and Jackson (1994). By using an assay that directly monitors the initiation step of translation, I found no effect on recognition of the first AUG codon when position +5 was varied in the series ACA, AAA, AGA, AUA (Table I, entries 5-8) or the series CCA, CAA, CGA, CUA (Figure 6C). The efficiency of initiation at AUG#1 was also affected very little when position +6 was varied in the series CCG, CCU, CCC, CCA (Table I, entries 9-12) or the series GCG, GCU, GCC, GCA (Table II, entries 1, 3, 5 and 7). Because I did not test all possible combinations, these experiments do not rule out the possibility that, as part of some particular sequence, A^{+5} and U^{+6} might be preferable to some alternative sequence (see below); but the experiments do preclude generalizing the optimal context for initiation to include A^{+5} and U^{+6}.

The experiments herein do suggest, on the other hand, that the positive effect of G^{+4} is generalizable. As illustrated in Figure 6, recognition of AUG#1 improved in response to six different codons that introduced G in position +4: GCG, GCU, GCC, GCA, GAU and GGA. A strong contribution of G^{+4} was also seen in the experiment in which a CUG codon was used in place of AUG#1 (Figure 7D). Speculation that the frequent occurrence of G in position +4 might simply reflect selection for Ala, Gly and Val as the penultimate amino acids (Flinta *et al.*, 1986), rather than a role for G^{+4} in initiation, no longer seems valid. The experiments herein, using an assay that directly monitors ribosome-mRNA initiation complexes, show that recognition of AUG start codons is stimulated strongly by G^{+4}.

augGU is not a favorable context for initiation

The one interesting situation in which G in position +4 failed to augment recognition of AUG#1 involves the sequence augGUN. As shown in Figure 6A, for example, augGUA (lane 11) was recognized with only the same low efficiency as augAUA (lane 12). At first glance, the data in Figure 6A (e.g. lane 9 versus lane 11) seem to confirm an earlier report that (gug)GAU is a much stronger initiation site than (gug)GUA (Boeck and Kolakofsky, 1994). That observation contributed to the idea that A^{+5} and U^{+6} might be part of the optimal context for initiation. However, the more extensive set of data in the present study shows it is not that A^{+5} and U^{+6} augment initiation, but that the usual stimulatory effect of

G^{+4} fails in the case of augGUA. Table II shows, for example, that augAAU (entry 10) is not recognized significantly better than augAUA (entry 12). Instead, Table II shows unexpectedly low recognition of AUG#1 when the flanking codon is GUA (entry 11) compared with every other mRNA that has G in position +4 (entries 1, 3, 5, 7 and 9).

This unexpected deficiency is not limited to the sequence augGUA. Figure 6B shows that the usual stimulatory effect of G^{+4} fails with every codon that begins with GU. The simplest interpretation is that the sequence GU in positions +4/+5 somehow distorts the mRNA and thus impairs AUG codon recognition by the scanning 40S ribosomal subunit.

Because the deleterious effect seems to be attributable to a particular flanking sequence (augGU) rather than to a particular flanking codon, it is not likely that the defect occurs after assembly of the 80S ribosome when a tRNA tries to enter the A site. It seems unlikely, for example, that augGUA is a poor initiation site because the complementary Val-tRNA is scarce (Zhang *et al.*, 1991) or because Val-tRNA is structurally incompatible with Met-$tRNA_i^{met}$ when both tRNAs line up on the 80S ribosome (Irwin *et al.*, 1995). Those explanations might be tenable if the defect were limited to augGUA; but augGUG, augGUU and augGUC were equally poor start sites. Experiments described herein specifically contradict the idea that scarcity of the elongator tRNA required to form the first peptide bond at AUG#1 might shift initiation to a downstream site. That hypothesis appears to be ruled out by the results shown in Figure 6, where augGCG (Figure 6A, lane 1) and augGGA (Figure 6B, lane 6) were recognized efficiently despite the low abundance in reticulocytes of tRNAs corresponding to GCG and GGA (Hatfield *et al.*, 1979, 1982).

Initiation is best studied with an initiation-limited assay

The initiation-limited assay used here obviates two problems that can confuse assessment of the degree to which leaky scanning, caused by mutations around AUG#1, allows access to AUG#2. One problem with standard protein synthesis assays is that elongating ribosomes advancing from the upstream start site can block access to a second start site downstream, thus making AUG#1 appear stronger (less leaky) than it really is (Fajardo and Shatkin, 1990; Kozak, 1995). This sort of distortion, called elongational occlusion, was avoided in the present study by using inhibitors that prevent ribosomes from advancing beyond the initiation step.

A second potential problem is that varying bases +4, +5 and +6, and hence varying the penultimate amino acid, might change the stability of the test protein. This was argued not to be relevant in other studies (Boeck and Kolakofsky, 1994; Grünert and Jackson, 1994) because the amino acid changes that would result from the mutations in positions +5 and +6 should not have rendered the polypeptide unstable according to the N-end rule. However, the beautifully elucidated N-end rule pathway for protein turnover applies to proteins derived by proteolytic processing (Bachmair *et al.*, 1986; de Groot *et al.*, 1991; Gonda *et al.*, 1989; Varshavsky, 1995). One should not necessarily expect the predictions of the N-end rule to apply to nascent polypeptides in which the subterminal amino acid is varied. Unlike proteolytically derived polypeptides, the N-terminus of nascent polypeptides is subject to modification by methionine aminopeptidases, acetyltransferases, N-terminal amidases and other enzymes that may affect protein stability (Moerschell *et al.*, 1990; Kendall and Bradshaw, 1992; Stewart *et al.*, 1994; Baker and Varshavsky, 1995). It is not known whether the extent of these N-terminal modifications varies among batches of reticulocyte lysate, or whether the high level synthesis of some proteins *in vitro* might exceed the capacity of endogenous modifying enzymes. Because of these uncertainties, it seems dangerous to assume that differences in protein accumulation in response to mutations in positions +4, +5 and +6 reflect an effect of these nucleotides on the initiation of translation.

The present study gets around this concern by replacing the customary protein accumulation assay with one that directly monitors the initiation step of translation. Indeed, had I relied simply on measurement of protein yields, I might have concluded that initiation at augAAA was 5-fold more efficient than at augAUA (Figure 2A, lanes 6 and 8). However, those two start sites functioned with identical efficiencies when initiation was assayed directly (Figure 5B; Table I, entries

6 and 8).

Non-AUG start sites have the same flanking sequence requirements as AUG start sites

There is no compelling evidence for *cis*-acting elements in mRNAs that act uniquely at CUG, ACG and GUG start sites. Instead, non-AUG start codons seem to show a stronger dependence on the same ancillary features that augment AUG codon recognition. In some studies, for example, mutation of A^{-3} nearly abolished initiation from a CUG or ACG codon (Peabody, 1987; Portis *et al.*, 1994). In the experiments described herein, initiation at a CUG codon was barely detectable in the absence of G^{+4} (Figure 7C and D). The enhancing effect of downstream secondary structure, previously demonstrated for AUG start sites (Kozak, 1990a), was also evident with CUG start sites in Figure 7B. The strong dependence on these ancillary features probably follows from the fact that alternative start codons can form only two, instead of the usual three, base pairs with the anticodon in eukaryotic Met-tRNA$_i^{met}$. Prokaryotes are similar in the sense that initiation at a weak UUG start site requires an unusually strong Shine-Dalgarno interaction (Weyens *et al.*, 1988).

In an earlier study, Boeck and Kolakofsky (1994) postulated that A^{+5} and U^{+6} specifically augment initiation at non-AUG start sites, but they did not include tests with AUG in place of GUG. A companion paper (Grünert and Jackson, 1994) reported, on the other hand, that the effects of mutating positions +5 and +6 around a CUG start codon were qualitatively similar to the effects at an AUG codon.

In the present study, most of the mutations were introduced around AUG codons because the poor initiation at non-AUG codons, even in the best of circumstances, makes quantitation difficult. However, the experiment shown in Figure 7D using a CUG start site confirms the conclusion reached for AUG start sites: that recognition of the initiator codon improves when G is substituted for A in position +4, while substitutions in position +5 have no discernible effect.

There is also no compelling evidence for *trans*-acting factors in eukaryotes that specifically recognize non-AUG start codons. Some interesting experiments in yeast in which certain mutations in eIF-2 were shown to activate a silent UUG codon (Donahue *et al.*, 1988; Dorris *et al.*, 1995) are sometimes cited as evidence for a UUG-specific initiation factor. However, augmented initiation at UUG could be explained if the mutations in eIF-2 simply enhance its non-specific binding to mRNA. This could allow erroneous initiation events (i.e. use of a codon that only partially matches the anticodon in Met-tRNA$_i^{met}$) in the same way that streptomycin induces errors during polypeptide elongation by strengthening non-specific contacts between ribosomes and tRNA, and thus decreasing dependence on specific codon-anticodon contacts.

Materials and methods

Construction of plasmids

Plasmids used herein were derived from Riboprobe vector pSP64 (Promega Corp.) into which a CAT cartridge (Pharmacia Biotech) was previously inserted at the *Bam*HI site (Kozak, 1989b). The vector had been modified previously by introducing a T7 phage promoter (Kozak, 1994) followed by the sequence GAAGCTAAAACAAATCAATCAATCAAAACAC<u>AAGCTT</u>. This synthetic 5′ non-coding sequence, which is devoid of secondary structure, was chosen because it supports efficient translation when an appropriate initiator codon is introduced downstream. Between the *Hin*dIII site (AAGCTT underlined above) and a nearby *Bam*HI site marked in Figure 1, I inserted synthetic deoxyoligonucleotides that contain two ATG (AUG) codons, as illustrated in Figure 1. Using the cassette mutagenesis technique depicted in Figure 1, I systematically varied the codon on the 3′ side of AUG#1. The plasmids and resulting mRNAs were named according to the sequence following the first start codon for translation: p-

augGCA, etc., in Figure 1; p-cugAGU, etc., in Figure 7.

Because the presence of secondary structure appropriately positioned downstream from an AUG or CUG codon can augment initiation (Kozak, 1990a), two different sequences were used downstream. Beginning at the *Bam*HI site marked in Figure 1, the sequence was either GAUCCAAAGUCAGCCAAAUCAA (oligonucleotide 8335) or GAUCCGGGUUCUCCCGGAUCAA (oligonucleotide 8336). The latter sequence can form a stem-loop structure with a stability of -19 kcal/mol (Kozak, 1990a). Constructs that contain oligonucleotide 8336 are identified explicitly in the text and figures. All mRNAs discussed without mentioning the downstream sequence contained the structure-free oligonucleotide 8335, as in the mRNAs depicted in Figure 1.

Standard recombinant techniques used for these constructions were described previously (Kozak, 1989b). Plasmids were propagated using *Escherichia coli* RR1 (Gibco/BRL). The structures of all plasmids were confirmed by appropriate dideoxy chain-termination sequencing reactions using Sequenase-2 (U.S. Biochemical Corp.).

Synthesis of capped mRNAs

CsCl-purified plasmid DNA, linearized by digestion with *Ava*I, was used as the template for transcription by T7 RNA polymerase. Transcription reactions were carried out at 37°C as described previously (Kozak, 1989b) except that, after a 12 min incubation with m7GpppG caps (10 U/ml, Pharmacia Biotech), the GTP concentration was increased to 500 μM and incubation was continued for another 60 min. The reactions contained RNase inhibitor (150 U/ml, Gibco/BRL).

To ensure uniformity, all transcripts intended for a given translation experiment were synthesized using aliquots from a common reaction mixture, which included a trace of [^{3}H]UTP to facilitate quantification. mRNAs were extracted with phenol and purified by application to pre-spun Sephadex G50 columns (Boehringer Mannheim).

Complete translation assay

For the standard protein accumulation assay, a rabbit reticulocyte translation system supplemented with [^{3}H]leucine (140 μCi/ml, sp. act. 180 Ci/mmol) was programed with mRNA (12 μg/ml) and incubated for 1 h at 30°C. The Flexi reticulocyte lysate (Promega Corp.), which constituted 50% of the final reaction volume, was supplemented with 60 mM KCl and 19 non-radioactive amino acids at 20 μM each. In addition to the endogenous Mg^{2+} (stated by the supplier for each batch of lysate), reactions were supplemented with $Mg(CH_3COO)_2$ to give a final Mg^{2+} concentration of 2 mM, unless otherwise stated in the text or figure legends. A standard Mg^{2+} concentration of 2 mM was chosen because it was shown previously to support a pattern of context-dependent initiation *in vitro* similar to what is seen *in vivo* (Kozak, 1989b). To minimize variation, aliquots from a common reaction mixture were used for translation of all mRNAs in a given experiment.

Translation products were analyzed by polyacrylamide gel electrophoresis as described previously (Kozak, 1989b). The gels were impregnated with Entensify (DuPont NEN) before autoradiography with Kodak X-omat AR film at -70°C.

Primer extension assay of initiation complexes

Prior to ribosome binding, each mRNA was annealed with a ^{32}P-labeled deoxyoligonucleotide that would serve to prime the final reverse transcriptase step. The 23 nucleotide primer CTCAAAATGTTCTTTACGATGCC is complementary to codons 16-23 in the CAT coding domain. The primer was first labeled at the 5′ end by incubation with T4 polynucleotide kinase and [γ-^{32}P]ATP (3000 Ci/mmol). An aliquot of the ^{32}P-labeled primer was then incubated with mRNA (~5 pmol of each) in 11 μl of 50 mM Tris-HCl (pH 7.5) for 2 min at 65°C followed by 10 min at 37°C. The primer-mRNA

complexes were transferred to wet ice and held briefly while the reticulocyte reaction mixtures were assembled.

A rabbit reticulocyte lysate was used under the conditions described above except that [^{3}H]leucine was omitted and inhibitors of elongation (sparsomycin at 200 μM and cycloheximide at 90 μg/ml) were added. These inhibitors cause accumulation of initiation complexes in which the 80S ribosome is held at the AUG codon. Aliquots of a common reaction mixture were dispensed to glass tubes which were pre-incubated for 2 min at 30°C before adding the mRNA-primer complexes. After 4 min incubation at 30°C to allow ribosomes to engage the mRNA, the samples were applied to Sepharose CL-4B columns (15×0.7 cm) at 4°C. The column elution buffer contained 50 mM Tris-HCl (pH 8.3), 40 mM KCl, 6 mM $MgCl_2$, 5 mM dithiothreitol (DTT) and cycloheximide at 90 μg/ml. Column purification was omitted when more than six mRNAs were tested at one time.

Sepharose column fractions that contained ^{32}P-labeled ribosome-mRNA complexes were supplemented with 600 μM dATP, dGTP, dCTP and dTTP and with murine leukemia virus reverse transcriptase (Gibco/BRL Superscript II, used at 2 U/μl). Incubation at 37°C for 15 min allowed the primer to be extended up to the position of the bound ribosome. The positions of ribosomes on each mRNA were deduced from the lengths of the primer extension products, as determined by co-electrophoresis with an RNA sequence ladder. Appropriate ladders were generated from dideoxy sequencing reactions carried out at 42°C with avian myeloblastosis virus reverse transcriptase. Denaturing 8% polyacrylamide gels were used for electrophoresis. Autoradiograms of the dried gels, obtained in most cases with Kodak LS film, were quantified by densitometry. When weak start codons were tested (e.g. CUG in Figure 7D), AR film was used with an intensifying screen at -70°C.

A previous study that used this primer extension (toeprinting) assay describes some additional details and controls (Kozak, 1995).

Acknowledgements

Research in my laboratory was supported by grant GM33915 from the National Institute of General Medical Sciences, National Institutes of Health.

References

- Bachmair, A. , Finley, D. and Varshavsky, A. (1986) *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science*, **234**, 179-186[Medline]
- Baker, R.T. and Varshavsky, A. (1995) Yeast N-terminal amidase. A new enzyme and component of the N-end rule pathway. *J. Biol. Chem.*, **270**, 12065-12074[Abstract/Full Text]
- Boeck, R. and Kolakofsky, D. (1994) Positions +5 and +6 can be major determinants of the efficiency of non-AUG initiation codons for protein synthesis. *EMBO J.*, **13**, 3608-3617[Abstract]
- de Groot, R.J. , Rümenapf, T. , Kuhn, R.J. , Strauss, E.G. and Strauss, J.H. (1991) Sindbis virus RNA polymerase is degraded by the N-end rule pathway. *Proc. Natl Acad. Sci. USA*, **88**, 8967-8971[Abstract]
- Donahue, T.F. , Cigan, A.M. , Pabich, E.K. and Valavicius, B.C. (1988) Mutations at a Zn(II) finger motif in the yeast eIF-2β gene alter ribosomal start-site selection during the scanning process. *Cell*, **54**, 621-632[Medline]
- Dorris, D.R. , Erickson, F.L. and Hannig, E.M. (1995) Mutations in *GCD11*, the structural gene for eIF-2γ in yeast, alter translational regulation of *GCN4* and the selection of the start site for protein synthesis. *EMBO J.*, **14**, 2239-2249[Abstract]
- Fajardo, J.E. and Shatkin, A.J. (1990) Translation of bicistronic viral mRNA in transfected cells: regulation at the level of elongation. *Proc. Natl Acad. Sci. USA*, **87**, 328-332[Abstract]
- Flinta, C. , Persson, B. , Jörnvall, H. and von Heijne, G. (1986) Sequence determinants of cytosolic N-terminal protein processing. *Eur. J. Biochem.*, **154**, 193-196[Abstract]

- Gonda, D.K. , Bachmair, A. , Wünning, I. , Tobias, J.W. , Lane, W.S. and Varshavsky, A. (1989) Universality and structure of the N-end rule. *J. Biol. Chem.*, **264**, 16700-16712[Abstract]
- Grünert, S. and Jackson, R.J. (1994) The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. *EMBO J.*, **13**, 3618-3630[Abstract]
- Hartz, D. , McPheeters, D.S. , Traut, R. and Gold, L. (1988) Extension inhibition analysis of translation initiation complexes. *Methods Enzymol.*, **164**, 419-425[Medline]
- Hatfield, D. , Matthews, C.R. and Rice, M. (1979) Aminoacyl-tRNA populations in mammalian cells. Chromatographic profiles and patterns of codon recognition. *Biochim. Biophys. Acta*, **564**, 414-423[Medline]
- Hatfield,D., Varricchio,F., Rice,M. and Forget,B.G. (1982) The aminoacyl-tRNA population of human reticulocytes. *J. Biol. Chem.*, **257**, 3183-3188.
- Irwin, B. , Heck, J.D. and Hatfield, G.W. (1995) Codon pair utilization biases influence translational elongation step times. *J. Biol. Chem.*, **270**, 22801-22806[Abstract/Full Text]
- Kendall, R.L. and Bradshaw, R.A. (1992) Isolation and characterization of the methionine aminopeptidase from porcine liver responsible for the co-translational processing of proteins. *J. Biol. Chem.*, **267**, 20667-20673 [Abstract]
- Kozak, M. (1986a) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283-292[Medline]
- Kozak, M. (1986b) Bifunctional messenger RNAs in eukaryotes. *Cell*, **47**, 481-483[Medline]
- Kozak, M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125-8148[Abstract]
- Kozak, M. (1989a) The scanning model for translation: an update. *J. Cell Biol.*, **108**, 229-241[Abstract]
- Kozak, M. (1989b) Context effects and (inefficient) initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.*, **9**, 5073-5080[Medline]
- Kozak, M. (1990a) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA*, **87**, 8301-8305[Abstract]
- Kozak, M. (1990b) Evaluation of the fidelity of initiation of translation in reticulocyte lysates from commercial sources. *Nucleic Acids Res.*, **18**, 2828[Medline]
- Kozak, M. (1991) An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell. Biol.*, **115**, 887-903[Abstract]
- Kozak, M. (1992) A consideration of alternative models for the initiation of translation in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.*, **27**, 385-402[Medline]
- Kozak, M. (1994) Features in the 5′ noncoding sequences of rabbit α and β-globin mRNAs that affect translational efficiency. *J. Mol. Biol.*, **235**, 95-110[Medline]
- Kozak, M. (1995) Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. *Proc. Natl Acad. Sci. USA*, **92**, 2662-2666[Abstract] (and p. 7134)
- Kozak, M. (1996) Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome*, **7**, 563-574[Medline]
- Kozak, M. and Shatkin, A.J. (1977) Sequences and properties of two ribosome binding sites from the small size class of reovirus messenger RNA. *J. Biol. Chem.*, **252**, 6895-6908[Medline]
- Lin, F.-T. , MacDougald, O.A. , Diehl, A.M. and Lane, M.D. (1993) A 30-kDa alternative translation product of the CCAAT/enhancer binding protein α message: transcriptional activator lacking antimitotic activity. *Proc. Natl Acad. Sci. USA*, **90**, 9606-9610[Abstract]
- Moerschell, R.P. , Hosokawa, Y. , Tsunasawa, S. and Sherman, F. (1990) The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo*. *J. Biol. Chem.*, **265**, 19638-19643[Abstract]
- Ossipow, V. , Descombes, P. and Schibler, U. (1993) CCAAT/enhancer binding protein mRNA is translated into multiple proteins with different transcription activation potentials. *Proc. Natl Acad. Sci. USA*, **90**, 8219-8223 [Abstract]
- Peabody, D.S. (1987) Translation initiation at an ACG triplet in mammalian cells. *J. Biol. Chem.*, **262**, 11847-11851[Abstract]
- Portis, J.L. , Spangrude, G.J. and McAtee, F.J. (1994) Identification of a sequence in the unique 5′ ORF of the gene encoding glycosylated gag which influences the incubation period of neurodegenerative disease induced by a murine retrovirus. *J. Virol.*, **68**, 3879-3887[Abstract]
- Stewart, A.E. , Arfin, S.M. and Bradshaw, R.A. (1994) Protein NH_2-terminal asparagine deamidase. *J. Biol. Chem.*, **269**, 23509-23517[Abstract]
- Varshavsky, A. (1995) The N-end rule. *Cold Spring Harbor Symp. Quant. Biol.*, **60**, 461-478[Medline]
- Weyens, G. , Charlier, D. , Roovers, M. , Piérard, A. and Glansdorff, N. (1988) On the role of the Shine-Dalgarno sequence in determining the efficiency of translation initiation at a weak start codon in the *car* operon of *Escherichia coli* K12. *J. Mol. Biol.*, **204**, 1045-1048[Medline]
- Zhang, S. , Zubay, G. and Goldman, E. (1991) Low-usage codons in *Escherichia coli*, yeast, fruit fly and primates.

. *Gene*, 105, 61-72[Medline]

Received on November 7, 1996; revised on January 13, 1997.

This article has been cited by other articles:

- Cao, W., Mattagajasingh, S. N., Xu, H., Kim, K., Fierlbeck, W., Deng, J., Lowenstein, C. J., Ballermann, B. J. (2002). TIMAP, a novel CAAX box protein regulated by TGF-beta 1 and expressed in endothelial cells. *Am. J. Physiol.* 283: C327-337 [Abstract] [Full Text]
- Sobczak, K., Krzyzosiak, W. J. (2002). Structural Determinants of BRCA1 Translational Regulation. *J. Biol. Chem.* 277: 17349-17358 [Abstract] [Full Text]
- Rodriguez, F., Harkins, S., Slifka, M. K., Whitton, J. L. (2002). Immunodominance in Virus-Induced CD8+ T-Cell Responses Is Dramatically Modified by DNA Immunization and Is Regulated by Gamma Interferon. *J. Virol.* 76: 4251-4259 [Abstract] [Full Text]
- Kakiuchi, M., Okino, N., Sueyoshi, N., Ichinose, S., Omori, A., Kawabata, S.-i., Yamaguchi, K., Ito, M. (2002). Purification, characterization, and cDNA cloning of {alpha}-N-acetylgalactosamine-specific lectin from starfish, Asterina pectinifera. *Glycobiology* 12: 85-94 [Abstract] [Full Text]
- Anelli, T., Alessio, M., Mezghrani, A., Simmen, T., Talamo, F., Bachi, A., Sitia, R. (2002). ERp44, a novel endoplasmic reticulum folding assistant of the thioredoxin family. *EMBO J.* 21: 835-844 [Abstract] [Full Text]
- CHANG, G.-J. J., DAVIS, B. S., HUNT, A. R., HOLMES, D. A., KUNO, G. (2001). Flavivirus DNA Vaccines: Current Status and Potential. *Annals NYAS Online* 951: 272-285 [Abstract] [Full Text]
- Kozak, M. (2001). Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* 29: 5226-5232 [Abstract] [Full Text]
- Kozak, M. (2001). A Progress Report on Translational Control in Eukaryotes. *Science's STKE* 2001: pe1-1 [Abstract] [Full Text]
- Nett, J. H., Kessl, J., Wenz, T., Trumpower, B. L. (2001). The AUG start codon of the Saccharomyces cerevisiae NFS1 gene can be substituted for by UUG without increased initiation of translation at downstream codons. *Eur J Biochem* 268: 5209-5214 [Abstract] [Full Text]
- Tu, Z., Ninos, J. M., Ma, Z., Wang, J.-W., Lemos, M. P., Desponts, C., Ghansah, T., Howson, J. M., Kerr, W. G. (2001). Embryonic and hematopoietic stem cells express a novel SH2-containing inositol 5'-phosphatase isoform that partners with the Grb2 adapter protein. *Blood* 98: 2028-2038 [Abstract] [Full Text]
- Hegde, M. R, Chong, B., Fawkner, M., Lambiris, N., Peters, H., Kenneson, A., Warren, S. T, Love, D. R, McGaughran, J. (2001). Microdeletion in the FMR-1 gene: an apparent null allele using routine clinical PCR amplification. *J. Med. Genet.* 38: 624-629 [Full Text]
- Gizard, F., Lavallee, B., DeWitte, F., Hum, D. W. (2001). A Novel Zinc Finger Protein TReP-132 Interacts with CBP/p300 to Regulate Human CYP11A1 Gene Expression. *J. Biol. Chem.* 276: 33881-33892 [Abstract] [Full Text]
- Sawant, S. V., Kiran, K., Singh, P. K., Tuli, R. (2001). Sequence Architecture Downstream of the Initiator Codon Enhances Gene Expression and Protein Stability in Plants. *Plant Physiol.* 126: 1630-1636 [Abstract] [Full Text]
- Rodriguez, F., Slifka, M. K., Harkins, S., Whitton, J. L. (2001). Two Overlapping Subdominant Epitopes Identified by DNA Immunization Induce Protective CD8+ T-Cell Populations with Differing Cytolytic Activities. *J. Virol.* 75: 7399-7409 [Abstract] [Full Text]
- Tailor, C. S., Marin, M., Nouri, A., Kavanaugh, M. P., Kabat, D. (2001). Truncated Forms of the Dual Function Human ASCT2 Neutral Amino Acid Transporter/Retroviral Receptor Are Translationally Initiated at Multiple Alternative CUG and GUG Codons. *J. Biol. Chem.* 276: 27221-27230 [Abstract] [Full Text]
- Tateishi, A., Inoue, H., Shiba, H., Yamaki, S. (2001). Molecular Cloning of {beta}-Galactosidase from Japanese Pear (Pyrus pyrifolia) and its Gene Expression with Fruit Ripening. *Plant Cell Physiol* 42: 492-498 [Abstract] [Full Text]
- Bonne, S, van Hengel, J, Nollet, F, Kools, P, van Roy, F (1999). Plakophilin-3, a novel armadillo-like protein present in nuclei and desmosomes of epithelial cells. *J Cell Sci* 112: 2265-2276 [Abstract]
- Pizzinat, N., Takesono, A., Lanier, S. M. (2001). Identification of a Truncated Form of the G-protein Regulator

- AGS3 in Heart That Lacks the Tetratricopeptide Repeat Domains. *J. Biol. Chem.* 276: 16601-16610 [Abstract] [Full Text]
- Jonassen, C. M., Jonassen, T. O., Saif, Y. M., Snodgrass, D. R., Ushijima, H., Shimizu, M., Grinde, B. (2001). Comparison of capsid sequences from human and animal astroviruses. *J Gen Virol* 82: 1061-1067 [Abstract] [Full Text]
- Rana, B. K, Shiina, T., Insel, P. A (2001). GENETIC VARIATIONS AND POLYMORPHISMS OF G PROTEIN-COUPLED RECEPTORS: Functional and Therapeutic Implications. *Annu. Rev. Pharmacol. Toxicol.* 41: 593-624 [Abstract] [Full Text]
- Weber, S., Fichtner, D., Mettenleiter, T. C., Mundt, E. (2001). Expression of VP5 of infectious pancreatic necrosis virus strain VR299 is initiated at the second in-frame start codon. *J Gen Virol* 82: 805-812 [Abstract] [Full Text]
- Dass, B., McMahon, K. W., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., MacDonald, C. C. (2001). The Gene for a Variant Form of the Polyadenylation Protein CstF-64 Is on Chromosome 19 and Is Expressed in Pachytene Spermatocytes in Mice. *J. Biol. Chem.* 276: 8044-8050 [Abstract] [Full Text]
- Cormont, M., Mari, M., Galmiche, A., Hofman, P., Le Marchand-Brustel, Y. (2001). A FYVE-finger-containing protein, Rabip4, is a Rab4 effector involved in early endosomal traffic. *Proc. Natl. Acad. Sci. U. S. A.* 98: 1637-1642 [Abstract] [Full Text]
- Guittaut, M., Charpentier, S., Normand, T., Dubois, M., Raimond, J., Legrand, A. (2001). Identification of an Internal Gene to the Human Galectin-3 Gene with Two Different Overlapping Reading Frames That Do Not Encode Galectin-3. *J. Biol. Chem.* 276: 2652-2657 [Abstract] [Full Text]
- Lopez-Huertas, E., Charlton, W. L., Johnson, B., Graham, I. A., Baker, A. (2000). Stress induces peroxisome biogenesis genes. *EMBO J.* 19: 6770-6777 [Abstract] [Full Text]
- Davuluri, R. V., Suzuki, Y., Sugano, S., Zhang, M. Q. (2000). CART Classification of Human 5' UTR Sequences. *Genome-Res.* 10: 1807-1816 [Abstract] [Full Text]
- Lohmann, J. U., Bosch, T. C.G. (2000). The novel peptide HEADY specifies apical fate in a simple radially symmetric metazoan. *Genes & Dev.* 14: 2771-2777 [Abstract] [Full Text]
- Barbier, O., Lévesque, E., Bélanger, A., Hum, D. W. (1999). UGT2B23, a Novel Uridine Diphosphate-Glucuronosyltransferase Enzyme Expressed in Steroid Target Tissues That Conjugates Androgen and Estrogen Metabolites. *Endocrinology* 140: 5538-5548 [Abstract] [Full Text]
- Doorbar, J., Elston, R. C., Napthine, S., Raj, K., Medcalf, E., Jackson, D., Coleman, N., Griffin, H. M., Masterson, P., Stacey, S., Mengistu, Y., Dunlop, J. (2000). The E1and E4 Protein of Human Papillomavirus Type 16 Associates with a Putative RNA Helicase through Sequences in Its C Terminus. *J. Virol.* 74: 10081-10095 [Abstract] [Full Text]
- Harborth, J., Weber, K., Osborn, M. (2000). GAS41, a Highly Conserved Protein in Eukaryotic Nuclei, Binds to NuMA. *J. Biol. Chem.* 275: 31979-31985 [Abstract] [Full Text]
- Gutiérrez-Escolano, A. L., Brito, Z. U., del Angel, R. M., Jiang, X. (2000). Interaction of Cellular Proteins with the 5' End of Norwalk Virus Genomic RNA. *J. Virol.* 74: 8558-8562 [Abstract] [Full Text]
- Grill, S., Gualerzi, C. O., Londei, P., Bläsi, U. (2000). Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. *EMBO J.* 19: 4101-4110 [Abstract] [Full Text]
- Yokogawa, T., Shimada, N., Takeuchi, N., Benkowski, L., Suzuki, T., Omori, A., Ueda, T., Nishikawa, K., Spremulli, L. L., Watanabe, K. (2000). Characterization and tRNA Recognition of Mammalian Mitochondrial Seryl-tRNA Synthetase. *J. Biol. Chem.* 275: 19913-19920 [Abstract] [Full Text]
- Nateri, A. S., Hughes, P. J., Stanway, G. (2000). In Vivo and In Vitro Identification of Structural and Sequence Elements of the Human Parechovirus 5' Untranslated Region Required for Internal Initiation. *J. Virol.* 74: 6269-6277 [Abstract] [Full Text]
- Lu, R., Misra, V. (2000). Zhangfei: a second cellular protein interacts with herpes simplex virus accessory factor HCF in a manner similar to Luman and VP16. *Nucleic Acids Res* 28: 2446-2454 [Abstract] [Full Text]
- Walikonis, R. S., Jensen, O. N., Mann, M., Provance, D. W. Jr, Mercer, J. A., Kennedy, M. B. (2000). Identification of Proteins in the Postsynaptic Density Fraction by Mass Spectrometry. *J. Neurosci.* 20: 4069-4080 [Abstract] [Full Text]
- Corcelette, S., Massé, T., Madjar, J.-J. (2000). Initiation of translation by non-AUG codons in human T-cell lymphotropic virus type I mRNA encoding both Rex and Tax regulatory proteins. *Nucleic Acids Res* 28: 1625-1634 [Abstract] [Full Text]
- Chang, G.-J. J., Hunt, A. R., Davis, B. (2000). A Single Intramuscular Injection of Recombinant Plasmid DNA Induces Protective Immunity and Prevents Japanese Encephalitis in Mice. *J. Virol.* 74: 4244-4252 [Abstract] [Full Text]
- Tujebajeva, R. M., Harney, J. W., Berry, M. J. (2000). Selenoprotein P Expression, Purification, and Immunochemical Characterization. *J. Biol. Chem.* 275: 6288-6294 [Abstract] [Full Text]
- Brewer, A., Gove, C., Davies, A., McNulty, C., Barrow, D., Koutsourakis, M., Farzaneh, F., Pizzey, J., Bomford, A., Patient, R. (1999). The Human and Mouse GATA-6 Genes Utilize Two Promoters and Two Initiation Codons. *J. Biol. Chem.* 274: 38004-38016 [Abstract] [Full Text]

- Lescure, A., Gautheret, D., Carbon, P., Krol, A. (1999). Novel Selenoproteins Identified in Silico and in Vivo by Using a Conserved RNA Structural Motif. *J. Biol. Chem.* 274: 38147-38154 [Abstract] [Full Text]
- Nakajima, T., Cheng, T., Rohrwasser, A., Bloem, L. J., Pratt, J. H., Inoue, I., Lalouel, J.-M. (1999). Functional Analysis of a Mutation Occurring between the Two In-frame AUG Codons of Human Angiotensinogen. *J. Biol. Chem.* 274: 35749-35755 [Abstract] [Full Text]
- Riechmann, J. L., Ito, T., Meyerowitz, E. M. (1999). Non-AUG Initiation of AGAMOUS mRNA Translation in Arabidopsis thaliana. *Mol. Cell. Biol.* 19: 8505-8512 [Abstract] [Full Text]
- Welch, E. M., Jacobson, A. (1999). An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. *EMBO J.* 18: 6134-6145 [Abstract] [Full Text]
- Gray, N. K., Wickens, M. (1998). CONTROL OF TRANSLATION INITIATION IN ANIMALS. *Annu. Rev. Cell Dev. Biol.* 14: 399-458 [Abstract] [Full Text]
- Van Eynde, A., Pérez-Callejón, E., Schoenmakers, E., Jacquemin, M., Stalmans, W., Bollen, M. (1999). Organization and alternate splice products of the gene encoding nuclear inhibitor of protein phosphatase-1 (NIPP-1). *Eur J Biochem* 261: 291-300 [Abstract] [Full Text]
- Kropotov, A., Sedova, V., Ivanov, V., Sazeeva, N., Tomilin, A., Krutilina, R., Oei, S. L., Griesenbeck, J., Buchlow, G., Tomilin, N. (1999). A novel human DNA-binding protein with sequence similarity to a subfamily of redox proteins which is able to repress RNA-polymerase-III-driven transcription of the Alu-family retroposons in vitro. *Eur J Biochem* 260: 336-346 [Abstract] [Full Text]
- Duga, S., Asselta, R., Del Giacco, L., Malcovati, M., Ronchi, S., Tenchini, M. L., Simonic, T. (1999). A new exon in the 5' untranslated region of the connexin32 gene. *Eur J Biochem* 259: 188-196 [Abstract] [Full Text]
- Bertilsson, G., Heidrich, J., Svensson, K., Asman, M., Jendeberg, L., Sydow-Backman, M., Ohlsson, R., Postlind, H., Blomquist, P., Berkenstam, A. (1998). Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc. Natl. Acad. Sci. U. S. A.* 95: 12208-12213 [Abstract] [Full Text]
- Dey, B. R., Spence, S. L., Nissley, P., Furlanetto, R. W. (1998). Interaction of Human Suppressor of Cytokine Signaling (SOCS)-2 with the Insulin-like Growth Factor-I Receptor. *J. Biol. Chem.* 273: 24095-24101 [Abstract] [Full Text]
- Drabkin, H. J., RajBhandary, U. L. (1998). Initiation of Protein Synthesis in Mammalian Cells with Codons Other Than AUG and Amino Acids Other Than Methionine. *Mol. Cell. Biol.* 18: 5140-5147 [Abstract] [Full Text]
- Naylor, D. J., Stines, A. P., Hoogenraad, N. J., Hoj, P. B. (1998). Evidence for the Existence of Distinct Mammalian Cytosolic, Microsomal, and Two Mitochondrial GrpE-like Proteins, the Co-chaperones of Specific Hsp70 Members. *J. Biol. Chem.* 273: 21169-21177 [Abstract] [Full Text]
- Fransen, M., Terlecky, S. R., Subramani, S. (1998). Identification of a human PTS1 receptor docking protein directly required for peroxisomal protein import. *Proc. Natl. Acad. Sci. U. S. A.* 95: 8087-8092 [Abstract] [Full Text]
- Bulbarelli, A., Valentini, A., DeSilvestris, M., Cappellini, M. D., Borgese, N. (1998). An Erythroid-Specific Transcript Generates the Soluble Form of NADH-Cytochrome b5 Reductase in Humans. *Blood* 92: 310-319 [Abstract] [Full Text]
- Chang, Y. E., Menotti, L., Filatov, F., Campadelli-Fiume, G., Roizman, B. (1998). UL27.5 Is a Novel gamma 2 Gene Antisense to the Herpes Simplex Virus 1 Gene Encoding Glycoprotein B. *J. Virol.* 72: 6056-6064 [Abstract] [Full Text]
- Smith, E. A., Fuchs, E. (1998). Defining the Interactions Between Intermediate Filaments and Desmosomes. *J. Cell Biol.* 141: 1229-1241 [Abstract] [Full Text]
- Rodriguez, F., An, L. L., Harkins, S., Zhang, J., Yokoyama, M., Widera, G., Fuller, J. T., Kincaid, C., Campbell, I. L., Whitton, J. L. (1998). DNA Immunization with Minigenes: Low Frequency of Memory Cytotoxic T Lymphocytes and Inefficient Antiviral Protection Are Rectified by Ubiquitination. *J. Virol.* 72: 5174-5181 [Abstract] [Full Text]
- Gladyshev, V. N., Jeang, K.-T., Wootton, J. C., Hatfield, D. L. (1998). A New Human Selenium-containing Protein. PURIFICATION, CHARACTERIZATION, AND cDNA SEQUENCE. *J. Biol. Chem.* 273: 8910-8915 [Abstract] [Full Text]
- Yu, J., Zhang, Y., McIlroy, J., Rordorf-Nikolic, T., Orr, G. A., Backer, J. M. (1998). Regulation of the p85/p110 Phosphatidylinositol 3'-Kinase: Stabilization and Inhibition of the p110alpha Catalytic Subunit by the p85 Regulatory Subunit. *Mol. Cell. Biol.* 18: 1379-1387 [Abstract] [Full Text]
- Passantino, R., Antona, V., Barbieri, G., Rubino, P., Melchionna, R., Cossu, G., Feo, S., Giallongo, A. (1998). Negative Regulation of beta Enolase Gene Transcription in Embryonic Muscle Is Dependent upon a Zinc Finger Factor That Binds to the G-rich Box within the Muscle-specific Enhancer. *J. Biol. Chem.* 273: 484-494 [Abstract] [Full Text]

Electrophoresis 1997, 18, 533–537

Leigh Anderson[1]
Jeff Seilhamer[2]

[1]Large Scale Biology Corporation, Rockville, MD, USA
[2]Incyte Pharmaceuticals, Palo Alto, CA, USA

A comparison of selected mRNA and protein abundances in human liver

In order to obtain an estimate of the overall level of correlation between mRNA and protein abundances for a well-characterized pharmaceutically relevant biological system, we have analyzed human liver by quantitative two-dimensional electrophoresis (for protein abundances) and by Transcript Image methodology (for mRNA abundances). Incyte's LifeSeq™ database was searched for expressed sequence tag (EST) sequences corresponding to a series of 23 proteins identified on 2-D maps in the Large Scale Biology (LSB) Molecular Anatomy™ database, resulting in estimated abundances for 19 messages (4 were undetected) among 7926 liver clones sequenced. A correlation coefficient of 0.48 was obtained between the mRNA and protein abundances determined by the two approaches, suggesting that post-transcriptional regulation of gene expression is a frequent phenomenon in higher organisms. A comparison with published data (Kawamoto, S., *et al.*, *Gene* 1996, *174*, 151–158) on the abundances of liver mRNAs for plasma proteins (secreted by the liver) suggests that higher abundance messages are strongly enriched in secreted sequences. Our data confirms this: of the 50 most abundant liver mRNAs, 29 coded for secreted proteins, while none of the 50 most abundant proteins appeared to be secreted products (although four plasma and red blood cell proteins were present in this group as contaminants from tissue blood).

1 Introduction

The control of gene expression is achieved by a series of complex mechanisms which can be divided into two basic phases. The first phase, which involves the processing of sequence information from DNA, through transcription, RNA splicing, and transport through the nuclear membrane to yield a mature mRNA, has been relatively well characterized for many genes through nucleic acid sequencing approaches. The second phase, involving translation into protein (dependent on mRNA translatability), folding, assembly into multimers, transport to an appropriate subcellular location, post-translational modifications, and final destruction, has been less comprehensively characterized. Both phases are likely to contain important control points associated with gene regulation underlying differentiation, disease processes and drug effects. For a variety of reasons, it would be useful to know the extent to which mRNA abundances are predictive of corresponding protein abundances. A series of powerful methodologies, including Transcript Imaging [1], SAGE [2], differential display [3] and array hybridization [4–6], have been developed to detect and in some cases quantitate differences in mRNA composition between different samples. In parallel, high resolution protein mapping systems, based on two-dimensional (2-D) electrophoresis [7], have been employed to build quantitative databases describing gene expression at the protein level [8–11]. By combining these approaches, it is possible for the first time to examine both levels at which gene expression is controlled, and thereby to develop a global understanding of gene expression control.

To date, we are aware of surprisingly little published work on the overall relationship of message and protein abundance, with the exception of a recent study by Kawamoto *et al.* [12], comparing mRNA levels obtained for plasma protein genes by transcript image methodology with the abundances of the corresponding plasma proteins in circulation. This report appeared to show a strong correlation between mRNA and protein abundance, based on data for nine human gene products. It seemed likely, however, that such secreted proteins constitute a special case, since they are rapidly delivered from the cell of synthesis to the plasma compartment, where many of the mechanisms that regulate cellular protein abundance are presumably absent. We therefore decided to compare mRNA and protein levels for a larger series of cellular molecules in order to see whether a simple relationship exists between mRNA and protein abundance for this class, and to see whether mRNAs for major cellular proteins are generally more or less abundant than those for major secreted products.

2 Materials and methods

Samples for 2-D electrophoresis were prepared by rapidly mixing a frozen powder of human liver (prepared and stored at liquid nitrogen temperature in the National Biomonitoring Specimen Bank at the US National Institute of Standards and Technology) with an 8-fold excess of 9 M urea, 2% NP-40, 1% mercaptoethanol and 2% carrier ampholytes (LKB 9–11). Ten µL of the resulting sample was analyzed using the Isodalt 2-D electrophoresis system, and the gels stained with colloidal Coomassie Brilliant Blue (CBB) G-250 as previously described [13–16]. Each stained slab gel was digitized in red light at 134 µm resolution using an Eikonix 1412 scanner and the digitized gel images pro-

Correspondence: Dr. Leigh Anderson, Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338 USA (Tel: +301-424-5989; Fax: +301-762-4892; email: leigh@lsbc.com)

Nonstandard abbreviation: CBB, Coomassie Brilliant Blue

Keywords: Messenger RNA / Two-dimensional polyacrylamide gel electrophoresis / Transcript image / Liver / Regulation

0173-0835/97/0304-0533 $17.50+.50/0

Table 1. Protein and mRNA abundances in human liver reported for 23 selected molecules

Protein name	Protein	Average protein abundance	Protein standard deviation	Average protein abundance (%)	Number of clones (BLAST)	Average message abundance
Carbamyl phosphate synthase	CPS	101475	12379	2.83	11	0.139%
Actin beta	ACTB	50345	17793	1.41	15	0.189%
Heat shock protein 60	HSP60	37656	1939	1.05	3	0.038%
Protein disulfide isomerase	PDI	31260	1942	0.87	2	0.025%
78 KD glucose regulated protein / BIP	BIP	31050	1993	0.87	1	0.013%
Calreticulin	CRTC	30491	2076	0.85	3	0.038%
F1 ATPase beta	F1ATPB	29539	1275	0.82	3	0.038%
Actin gamma	ACTG	23316	9012	0.65	17	0.215%
Heat shock cognate 70	HSC70	21647	908	0.60	1	0.013%
Cytochrome B5	CYB5	18776	1656	0.52	7	0.088%
Endoplasmin	ENPL	17817	5829	0.50	5	0.063%
75 KD glucose regulated protein	GR75	16380	1821	0.46	1	0.013%
Pyruvate carboxylase	PYVC	14655	1930	0.41	0	Not detected
Heat shock protein 70	HSP70	8629	1565	0.24	1	0.013%
Tubulin beta 1	TBB1	7125	1472	0.20	3	0.038%
Vimentin	VIME	6269	952	0.18	0	not detected
Tropomyosin	TPM	4090	600	0.11	1	0.013%
NADPH cytochrome P-450 reductase	NP450R	3303	1319	0.09	0	Not detected
Tubulin alpha 1	TBA1	3097	1409	0.09	5	0.063%
Heat shock protein 90	HSP90	2740	597	0.08	2	0.025%
Cytochrome oxidase II (mit encoded)	COX-II	2384	651	0.07	0	Not measured
Laminin receptor	LAMR	1531	602	0.04	4	0.050%
Lumin B	LAMB	1454	371	0.04	2	0.025%

a) Protein abundance is given in pixel-gray levels (the integrated CBB optical density of the appropriate spot or spots on a 2-D gel), where multiple spots comprising a single gene product have been summed. Messenger RNA measurements are given as a percentage of the total number of clones sequenced in the relevant transcript images.

cessed using the Kepler® software system (Large Scale Biology) to give protein abundances in terms of pixel × gray-level values, as well as group average abundances and standard deviations over a set of seven male human livers. Relative abundances were computed by dividing individual average abundances by the average total abundance of the proteins resolved on the gels. A series of proteins was identified on these gels based on close homology with identified rodent liver spots and on identifications published by Hughes *et al.* [17]. Total cellular RNA was extracted from samples of human liver tissue by the method of Chirgwin *et al.* [18], and poly-A+ RNA was prepared by hybridization to oligo-dT cellulose. Five μg of poly-A+ RNA was used to construct a cDNA library using the Gubler and Hoffman method [19] in bacteriophage-lambda UNIZap™ (Stratagene Inc., La Jolla, CA). The library was converted to plasmid DNA by bulk excision, and individual colonies were selected for DNA template preps. The templates were sequenced enzymatically (Sanger *et al.* [20]) on an ABI 373 automated DNA sequencer. Templates considered sequenced sucessfully contained > 230 bases of cDNA insert sequence after removal of repetitive and low information sequences, > 90% base call accuracy, and were not of mitochondrial, vector or host origin. Resulting DNA sequences were analyzed using the BLAST program for similarity with other known primate, mammalian, and subsequently all divisions of GenBank. Similarity data was stored and tabulated in the LifeSeq™ software (Incyte, Palo Alto, CA), from which relative fractions of specific gene products present within the starting RNA prep were calculated as follows: % abundance = # clones representing each gene / total # of genes sampled *100. A total of 7925 clones were sequenced from liver obtained from two individuals: one male (5054 clones) and one female (2871 clones). Data from Table 1 of Kawamoto *et al.* [12], was replotted using protein abundances for human plasma proteins taken as mean values of the range presented in reference [21]. An error in the abundance of the haptoglobin α1s polypeptide (which was assumed in [12] to account for the entire abundance of the haptoglobin $\alpha_1\beta_1$ tetramer) was corrected.

3 Results

Protein and mRNA abundance data were collected for a set of gene products identified on 2-D gels (Table 1). Standard deviations of the protein measurements across six individual livers were relatively low, averaging 19% of the mean abundance. Of the 23 selected proteins, mRNAs for 19 were detected in human liver transcript images. Of these 19, five were represented by 1 clone, three by 2 clones, four by 3 clones, and the rest by between 4 and 17 clones. Of the four gene products undetected at the mRNA level, one (cytochrome oxidase subunit II: COX-II) was deleted from the Transript Image dataset during standard initial sequence data workup, which removes all mitochondrial sequences. A plot of protein abundance (expressed as integrated Coomassie Blue absorbance averaged over seven individual livers) *versus* mRNA abundance (expressed as per-

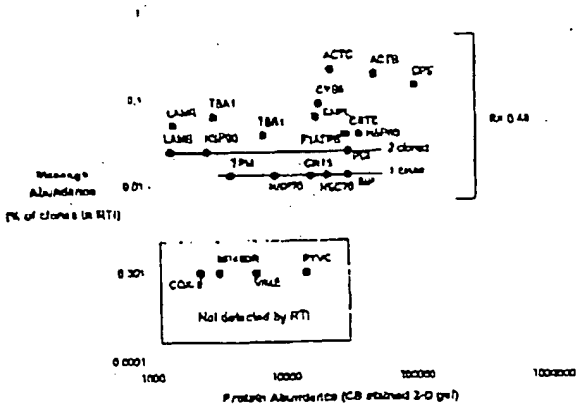


Figure 1. A log-log plot of the abundances of each of 23 gene products at the protein level (*X*-axis) and mRNA level (*Y*-axis). Four proteins for which mRNA measurements were not available – three for which no clones were detected, and one intentionally deleted from the RTI dataset (COX-II) – are shown boxed at the lower left, with correct relative protein abundances. The Pearson product moment correlation coefficient between the two sets of 19 valid measurements is 0.48. Each measurement is labeled with a code whose identity is shown in Table 1.

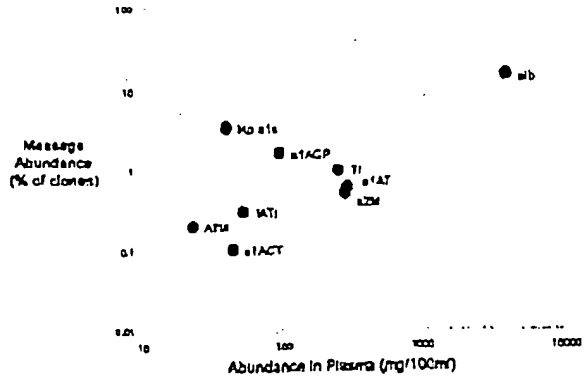


Figure 2. A log-log plot of data on mRNA abundance taken from Kawamoto *et al.* [12] *versus* average protein abundances in plasma taken from [21]. The protein abundance value for the haptoglobin α1s polypeptide has been corrected to reflect the fact that this subunit accounts for only 21% of the mass of the haptoglobin $\alpha_2\beta_2$ tetramer.

centage of total cDNA clones in the transcript images of two livers) indicates a modest correlation between the two (Fig. 1). The Pearson product-moment correlation coefficient obtained from the 19 pairs of measurements is 0.48. The abundance values obtained at the protein level spanned a 70-fold range, while the detectable mRNA abundances spanned a 16-fold range for these genes (although the latter value may reflect the limited number of clones sequenced). One particularly interesting subset of measurements concerns the β and γ actins. Here the mRNA abundances are, respectively, 0.189% and 0.215%, whereas the protein abundances are, respectively, 1.41% and 0.65% of the total. In this comparison, both sets of measurements are likely to be quite accurate, since numerous clones were detected for each of the two messages, and since the two proteins are so homologous, and have such close p*I*s, that they should bind CBB similarly. Nevertheless, the relative abundances at the RNA and protein levels are inverted (β actin is the more abundant protein, while γ actin has the more abundant message), and the mRNA:protein ratios for the two genes differ by more than a factor of two. Carbamyl phosphate synthase (CPS), the most abundant protein detected in liver over the p*I* range of conventional 2-D gels (pH 4–7), had a relative abundance of 2.83% (protein) and yet comprised only 0.139% of the total message (less than either actin). In this case, the mature protein is sequestered inside the mitochondrion, and therefore might be expected to show slow turnover and a consequent large disparity between mRNA and protein abundance.

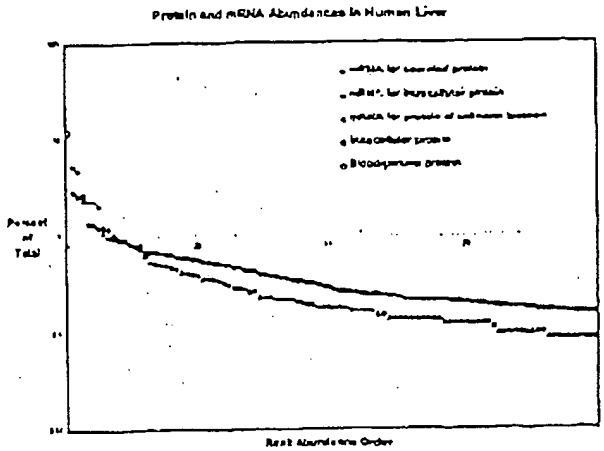


Figure 3. Relative abundance distributions of the top-ranked 100 mRNAs and proteins detected in human liver. The first (leftmost) molecule is the most abundant, followed by molecules of decreasing abundance through the 100th rank (at the right). Abundances of both mRNAs and proteins are plotted as a percentage of total detected molecules on a log scale. Message and protein points at the same rank are not, in general, products of the same gene.

A reexamination (Fig. 2) of the data of [12] on genes for plasma proteins, using estimates for corresponding protein abundances revised to account for the $\alpha_2\beta_2$ structure of haptoglobin, showed a higher correlation coefficient between mRNA and protein abundance (0.96). This value is probably exaggerated due to the large separation of the albumin values from the rest of the data: if albumin is omitted from the calculation, the correlation coefficient drops to −0.19. However, it is clear that the plasma proteins are represented by many more mRNA copies than major cellular proteins: albumin, for example, accounts for about 14% of the total number of clones examined [12], with a number of other plasma proteins accounting for more than 1% of the total each. By contrast, none of the cellular proteins chosen from the 2-D gel data accounted for much more than 0.1% of the mRNAs sequenced. To further pursue this observation, we compared the relative abundance distributions of the 100 top-ranked (most abundant) mRNAs and proteins in our data sets (Fig. 3). Forty-one of the top 100 mRNAs, and 29 of the top 50, coded for proteins known, or expected, from sequence to be secreted from the liver, while none of the top 100 proteins appeared to be secretory forms of the human plasma proteins. The two most

abundant proteins in these samples (hemoglobin β and albumin) as well as two of lower abundance (α1 antiprotease and transferrin) were blood proteins that constitute contaminants of the liver in this context-proteins which would have been removed by perfusion.

4 Discussion

Despite extensive work on the regulation of many individual genes, little attention appears to have been paid to the global question of the relation between mRNA and corresponding protein abundance in eukaryotes. We have attempted to provide an initial estimate of the relationship of mRNA and corresponding cellular protein abundances through use of correspondences between two databases: the Molecular Anatomy™ (2-D gel) and LifeSeq™ (Transcript Image) databases of human liver. Using a panel of 23 proteins identified on 2-D gels of human liver, we searched LifeSeq™ to determine the number of clones matching the corresponding gene sequence by BLAST. Matches were found for 19 proteins, and the correlation coefficient obtained over this set of data was 0.48. This number is intriguingly close to the middle position between a perfect correlation (1.0) and no correlation whatever (0.0). One simple interpretation of such a value is that the two major phases of gene expression regulation (transcription through message degradation on the one hand, and translation through protein degradation on the other) are of approximately equal importance in determining the net output of functional gene product (protein). Several issues may limit the quantitative accuracy of this result. First, the protein measurements rely on CBB binding to a series of different proteins. Although the measurements obtained show good (low) standard deviations across a set of six individual livers, it is well known that different proteins can bind CBB with different affinities. Thus the measurement scale for one protein may differ from another by up to approximately twofold. Since, however, these relative scale errors should be normally distributed, we expect them to have little effect on the overall correlation. Precision of the mRNA measurements is also limited, in this case because a limited number of clones was detected for the selected proteins. Five genes, for example, were represented by only one clone each among the 7925 clones sequenced from the respective cDNA tissue libraries. This low relative expression at the mRNA level is expected, since a majority of the high abundance mRNAs in liver code for plasma proteins. However, such small numbers of clones lead to potentially large quantitative errors because of sampling error. Here again, we believe these errors should be relatively random across the set of proteins chosen, and thus should not skew the result appreciably. A third potential difficulty is that the databases used for the protein and mRNA abundance estimates were prepared from different samples. In future, it will thus be of great interest to repeat the experiment using the same samples to examine both mRNA and protein abundances.

Despite these potential sources of error, at least one homologous pair of proteins (the β and γ actins) shows persuasive evidence of post-transcriptional regulation, with mRNA-to-protein ratios differing by more than a factor of two between the two genes. This is a particularly striking case since the two proteins are essentially indistinguishable in function (apart from affinitiy for MgADP; 22), have very similar sequences, and are produced in a constant ratio (approximately 2:1 in males) in virtually all cell types. One possible alternative explanation could be a sex difference in liver expression of γ actin, as is seen in rodents [23] where γ actin protein expression averages almost twice as high in females as males. This seems unlikely since 64% of clones in the RTI data were from male liver, and all the 2-D data was from male livers.

An analogous set of data for plasma proteins secreted by the liver has been published by Kawamoto *et al.* [12] and we have reanalyzed their values to see whether a similar mRNA-to-protein relationship holds. It appears, based on nine plasma proteins, that a higher correlation coefficient applies: 0.96. This result is less convincing, however, because one gene product (albumin) is well-separated from the cluster of the remaining eight, and thus exercises a disproportionate influence on the correlation coefficient. In fact, if albumin is omitted from the calculation, the correlation coefficient is reduced to −0.19, which suggests a very poor correlation.

What is perhaps more striking is the relatively much higher abundance of the plasma protein mRNAs as compared to major cellular proteins such as carbamyl phosphate synthase, the actins, or cytochrome b5. Mid-abundance plasma proteins were represented by mRNAs having approximately 100-fold higher relative abundance than mid-abundance cellular proteins. This result is verified by a direct comparison of the relative abundance distributions of the 100 top-ranked mRNAs and proteins in our data sets (which are, in general, different sets of genes). Twenty-nine of the top 50 messages are secreted products, while none of the top 50 proteins appear to be the pro- form of a secreted molecule. Such a conclusion is not surprising, since the liver is responsible for generating high protein concentrations in the relatively large plasma compartment of the body, but does so by means of closely coupled synthesis and secretion with little accumulation of precursor proteins in process. This points to a potentially significant difference in the pictures obtained from mRNA and protein abundance databases. Major secreted proteins appear to have much more abundant mRNAs than many important cellular proteins, and hence mRNA abundance databases that concentrate on a small number of the highest abundance messages may be biased towards secreted proteins over cellular molecules. This represents an advantage of the mRNA approach relative to protein databases in the search for novel cytokines and other secreted proteins, but a disadvantage in the characterization of cellular metabolic and control processes. Additionally, it suggests that mRNAs for secreted proteins may have, on the whole, shorter half-lives than mRNAs for cellular enzymes, the latter being more frequently regulated at the translational level.

We also found important differences in the overall shapes of the relative abundance distributions of the 100

top-ranked mRNAs and proteins. While both distributions contain a few very high abundance molecules (in the 3–10% range) they appear to diverge significantly below the 15th most abundant gene product, with proteins 16–100 accounting for roughly twice as high a relative abundance as the 16th–100th mRNAs. Not all proteins are represented on the 2-D gels used here (which fail to resolve proteins with $pI > 7$), but the estimated 40% of proteins thus excluded would not affect the shape of the distribution over positions 50–100 significantly if they have an abundance distribution similar to the pI 4–7 proteins (based on a simulation using the data shown). The mRNA abundance distribution covers all cloned messages (not a subset of genes), and for abundant mRNAs it should be complete as it stands. Altogether, the top 100 mRNAs comprise 51.3% of the total clones, while the top 100 proteins comprise 63.1% of the total protein detected. Hence it appears likely that the distribution of protein abundances is significantly different from that of mRNAs, showing a more gradual falloff in the region examined, and that techniques able to detect down to a specified percent abundance threshold would reveal more proteins at a given threshold than mRNAs. As the protein and nucleic acid databases expand, we anticipate the possibility of generating successively more robust estimates of the global relationship between mRNA and protein abundance, and thus a better understanding of multi-level gene expression control in complex organisms such as man.

Human liver samples analyzed by 2-D electrophoresis were kindly provided by the National Biomonitoring Specimen Bank at the US National Institute of Standards and Technology under the direction of Dr. Stephen Wise.

Received November 27, 1996

5 References

[1] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Matsubara, F. A., *Nat. Genet.* 1992, *2*, 173–179.
[2] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, *270*, 484–487.
[3] Liang, P., Pardee, A. B., *Curr. Opin. Immunol.* 1995, *7*, 274–280.
[4] Augenlicht, L. H., Wahrman, M. Z., Halsey, H., Anderson, L., Taylor, J., Lipkin, M., *Cancer Res.* 1987, *47*, 6017–6021.
[5] Fodor, S. P., Rava, R. P., Huang, X. C., Pease, A. C., Holmes, C. P., Adams, C. L., *Nature* 1993, *364*, 555–556.
[6] Schena, M., Shalon, D., Davis, R. W., Brown, P. O., *Science* 1995, *270*, 467–470.
[7] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.
[8] Garrels, J. I., Futcher, B., Kobayashi, R., Latter, G. I., Schwender, B., Volpe, T., Warner, J. R., McLaughlin, C. S., *Electrophoresis* 1994, *15*, 1466–1486.
[9] Celis, J. E., Rasmussen, H. H., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Gromov, P., Vorum, H., Vessilev, A., Baskin, Y., Liu, X., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A. H., Wulbum, E., Kjærgaard, A., Andersen, S., Puype, M., Van Damme, J., Vanderkerckove, J., *Electrophoresis* 1994, *15*, 1349–1458.
[10] Hochstrasser, D. F., Frutiger, S., Paquet, N., Bairoch, A., Ravier, F., Pasquali, C., Sanchez, J. C., Tissot, J. D., Bjellqvist, B., Vargas, R., Appel, R. D., Hughes, G. J., *Electrophoresis* 1992, *13*, 992–1001.
[11] Anderson, N. L., Esquer-Blasco, R., Hofmann, J. P., Meheus, L., Raymackers, J., Steiner, S., Witzmann, F., Anderson, N. G., *Electrophoresis* 1995, *16*, 1977–1981.
[12] Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K., Matsubara, K., *Gene* 1996, *174*, 151–158.
[13] Anderson, N. L., Anderson, N. G., *Anal. Biochem.* 1978, *85*, 341–354.
[14] Anderson, N. G., Anderson, N. L., *Anal. Biochem.* 1978, *85*, 331–340.
[15] Anderson, N. L., Esquer-Blasco, R., Hofmann, J.-P., Anderson, N. G., *Electrophoresis* 1991, *12*, 907–930.
[16] Anderson, N. L., Large Scale Biology Press, Washington, DC 1991, ISBN 0-945532-01-6, 200 pp., and http://www.lsbc.com.
[17] Hughes, G. J., Frutiger, S., Paquet, N., Pasquali, C., Sanchez, J.-C., Tissot, J.-D., Bairoch, A., Appel, R., Hochstrasser, D., *Electrophoresis* 1993, *14*, 1216–1222.
[18] Chirgwin, J., Przybla, A., MacDonald, R., Rutter, W., *Biochemistry* 1979, *18*, 5294–5299.
[19] Gubler, U., Hoffman, B. J., *Gene* 1983, *25*, 263–269.
[20] Sanger, F., Nicklen, S., Coulson, A. R., *Proc. Nat. Acad. Sci. USA* 1977, *74*, 5463–5469.
[21] Putnam, F. W., (Ed.), *The plasma proteins*, Academic Press, New York 1975, pp. 26–29.
[22] Anderson, N. L., *Biochem. Biophys. Res. Comm.* 1979, *89*, 486–490.
[23] Anderson, N. L., Giere, F. A., Nance, S. L., Gemmell, M. A., Tollaksen, S. L., Anderson, N. G., Galteau, M.-M., Siest, G. (Eds.), *Progrès Récents en Electrophorise Bidimensionelle*, Presses Universitaires de Nancy, Nancy 1986, pp. 253–260.

Jane M. C. Oh
Franck Brichory
Eric Puravs
Rork Kuick
Chris Wood
Jean Marie Rouillard
John Tra
Sharon Kardia
David Beer
Samir Hanash

University of Michigan
Medical Center,
Ann Arbor, MI, USA

A database of protein expression in lung cancer

We have developed a comprehensive approach to identifying molecular changes in lung cancer that includes both genomic and proteomic analyses. The related effort has produced a large amount of data pertaining to gene expression at the RNA and protein levels. As a result, we have constructed a database that contains protein expression data on lung cancer as well as other relevant data including DNA microarray derived data. A large number of proteins that are expressed in different types of lung cancer have been identified and have been correlated with the expression measures for their corresponding genes at the RNA level. The database is intended to facilitate our effort at developing novel classification schemes for lung cancer and the identification of novel markers for early diagnosis.

Keywords: Lung / Cancer / Database / Microarray PRO 0131

1 Introduction

There is substantial interest in implementing novel and comprehensive strategies for the molecular analysis of tumors and relevant biological fluids. We have implemented a strategy for the molecular analysis of lung cancer that integrates genomic analysis using genome scanning procedures, transcriptomic analysis using cDNA and oligonucleotide microarrays, and proteomic analysis. For the latter, we have relied to date primarily on 2-D polyacrylamide gels. However the 2-D gel approach is being increasingly complemented with additional analyses using liquid based protein separations and protein microarrays. While on the one hand proteomic analysis complements genomic analysis for a global assessment of gene expression, on the other hand proteomic analysis uniquely contributes an understanding of protein post-translational modifications and the location of protein gene products in subcellular compartments. The scope of our overall molecular analysis study of lung cancer is shown in Fig. 1. Important objectives include the development of novel molecular classification schemes for lung cancer and the identification of novel markers for the early detection of lung cancer.

The large body of proteomic and other data we have collected has necessitated the construction of a database in which basic and derived data is organized. There have been relevant related efforts at databasing of 2-D data by other groups. One such database is the 2DWG Meta-database of 2-D gel images, which contains 2-D derived data acquired by a combination of review of results as well as submissions by investigators [1]. However, to date there are only three entries found matching the query for human lung images in the 2DWG Web Gel Meta-database web site (http://www-lecb.ncifcrf.gov/2dwgDB). The database we have constructed, in its entirety, is relevant to a variety of cancers. However the focus of this review is the use of the database to achieve our objectives related to the molecular analysis of lung cancer specifically. The goal of the database is to facilitate planned analyses, *i.e.* statistical analysis, as well as post-planned analyses, *i.e.* data mining. The intent is to make the database queryable on a protein – by – protein basis as well as through subgrouping of samples analyzed, in a menu driven fashion. Internet and WWW technologies are used not only to allow investigators to view visual and textual data together, but also to allow investigators in other locations to retrieve archival data using different computer systems.

2 Laboratory information processing system

A long-standing Laboratory information processing system (LIPS) developed by our group [2] has been adapted for our database. LIPS consists of multiple systems and processes. A variety of data is stored in a variety of formats with individualized programs for viewing the data. Typical processes using LIPS include: sample inventory; digitize images; detect and quantify spots; match spots and normalize spot sizes across images, choose spots for MS analysis, enter profiles from MS-Fit web search; transfer data to statistical software or spreadsheets.

Data tend to be complex and dynamic in that their contents are ever changing as information is added, modified or removed. Simple or intensive analyses of 2-D patterns

Correspondence: Dr. S. Hanash, University of Michigan Medical Center, 1150 W. Medical Center Drive, A520 Medical Science Research Building I, Ann Arbor MI 48109–0656, USA
E-mail: shanash@umich.edu
Fax: +1-734-647-8148

Abbreviation: LTPS, Laboratory information processing system

1615-9853/01/1009–1303 $17.50+.50/0

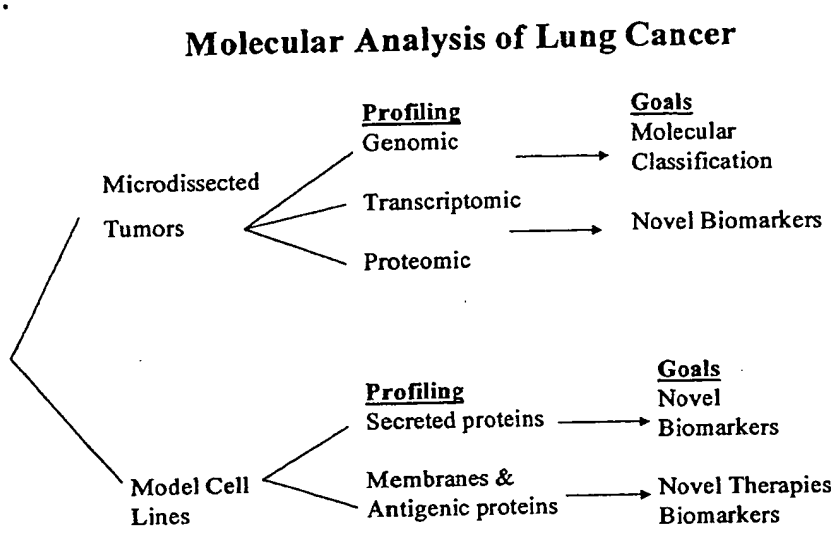


Figure 1. Methods and goals of lung cancer studies.

have produced a large amount of data. Data is both textual (*e.g.*, reports and numbers) and visual (*e.g.*, 1-D and 2-D gel images).

Some types of data generated by LIPS include: 2-D protein gel images (silver, modified silver, blots, ^{35}S labeled gels); genome scans; 1-D gel images; spot information-protein names; gene information from DNA microarrays; MS files and MS-Fit reports (Word documents); figures (Raster files on the Sun and actual photographs); data from protein microarrays; data from liquid chromatography separations.

However, as computer technology has evolved, quantum jumps in improvements in organizing unstructured, scientific data into a structured database have become possible. A major function of our database and its interfaces is to serve as a computer-based tool for capturing the basic quantitative data from 2-D gel images and derived data and findings derived from different studies about proteins detected in 2-D patterns of various tumor types [3]. As a result, investigators are provided with easy access to data as well as a means for intelligent data mining of the existing data. A logical view of the database schema is shown in Fig. 2 and a list of tables and their attributes are shown in Table 1.

The following are important features of the 2-D gel related component of our lung protein database:

(1) All 2-D gel images are placed in hierarchies so that: (a) every study image is matched to one master image, *i.e.* all lung adenocarcinoma tumor images are matched to one master image; (b) every master image is matched to at most one (higher) master image, *i.e.* all masters for different lung tumor types are matched to one tumor master. This allows the database to have an indexing mechanism that can relate a spot to any gel in the hierarchy. The database provides a capability to access the basic and derived data using the following types of queries: (a) given a spot on any gel, find all spots that are matched to it; (b) given a spot on any gel, find all protein identifications made for it, and (c) given a spot on any gel, find all findings/conclusions that are linked to it.

(2) All samples (and thereby gels derived from them) are identified by a list of source characteristics in four major categories: experiment code; cell type code; treatment code; and fraction code. This allows the database to have an identification mechanism that can relate a gel to any source in the hierarchy. The database provides a capability to find all images as follows: (a) given a category, find all images that have the same value of the category; and (b) given any combination of four categories, find all images that satisfy the condition.

(3) All protein spots are identified by a list of characteristics in four major attributes: protein name; p*I* and M_r; accession number; and protein sequence data. A spot may have several findings and there may be many kinds of findings derived from a particular study. If possible the findings are recorded in a consistent way, however this is not always possible due to some characteristics of such findings (*e.g.*, statistical analysis matrices, MS data, and Affymetrix data). As the number of studies has increased, the amount of data produced has increased. Some of the data (*e.g.* mass spectra and Affymetrix (Santa Clara, CA, USA) oligonucleotide chip readouts) is very large, and fills up the hard disks of the computers where it is collected. Such data is generally saved on CD-Rs, and only the most recent data is kept in a computer. It is sometimes easier

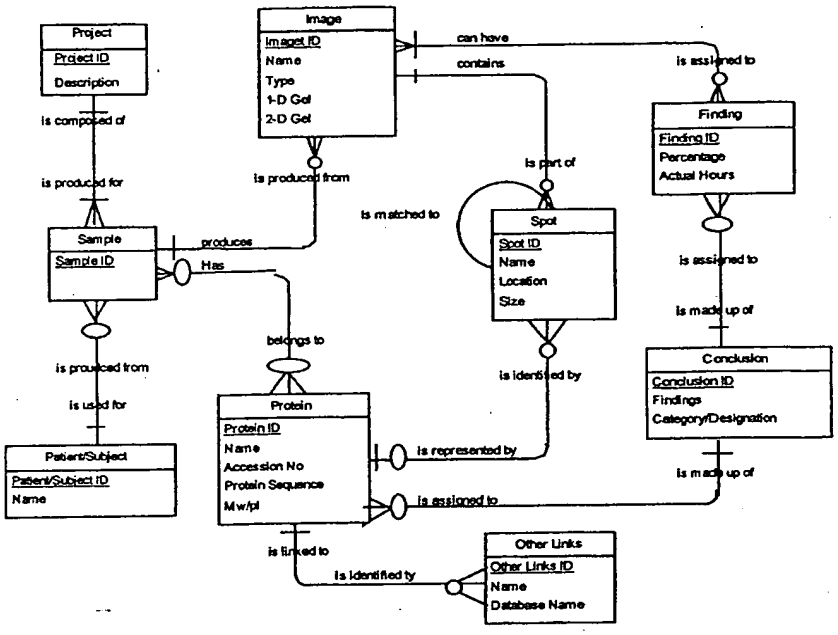


Figure 2. A logical view of database schema.

Table 1. A list of tables and their attributes in the lung protein database

Table name	Unique identifier (Primary Key)	List of attribute types
Project	Project Name	Project Type, Description
Sub Project	Sub Project Name	Date Started, Comment
Subject	Subject ID	Case No, Sex, Birthdate, Comment
Tissue Sample	Tissue Sample ID	Tissue Type, Diagnosis, Date Sample Taken, Date Received, How Received, Source, Comment
DNA Sample	DNA Sample ID	Date Produced, Concentration, Freezer Location, Comment
Gel	Gel ID	Sample ID, Batch ID, Enzyme Combination, Electrophoresis Process, Comment
Image	Image Name	Date Imaged, Exposure Time, Image Type, Image Location, Comment
Spot	Image Name & Spot No	X, Y, Intensity, Spot Type
Match	Match ID	Master Image Name, Master Spot No, Image Name, Spot No
Experiment	Experiment Code	Description
Cell Type	Cell Code	Description
Treatment	Treatment Code	Description
Fraction	Fraction Code	Description
Protein Sample	Sample ID	Experiment Code, Cell Code, Treatment Code, Treatment Date, Fraction Code, Comment, Project Type, Gel ID, Image Name, Image Type, Researcher
Protein	Protein Name	Image Name, Spot No
Other Link	Protein Link ID	Protein Name, Database Name, URL
Findings	Image Name & Spot No	Category, Designation, Finding
Protein Identification	Image Name & Spot No	Accession No, cDNA cloning, Cell Lines, Facility, Date, Genomic Cloning, Glycosylation, M_r, pI, Phosphorylation, Phosphorylation Residues, Related Spot, Sequences, Source of Protein, Name, Structural Modification, Subcellular Localization, Tissue Distribution, Type of Membrane, Type of Sequencing

to post individual files on the web. Individual web pages have been created with textual and visual data that are difficult to relate in a table. This allows investigators an opportunity to analyze 2-D gel and other images containing spots that have not been detected or identified and to compare data across studies. In addition this is used to link our data to other biological knowledge repositories such as GenBank, PIR International, and SWISS-PROT.

3 Contents of the lung cancer protein database

A large number of studies involving lung cancer have been independently performed in the laboratory. At the protein level, these studies have resulted in 1349 images, over 1000 of which are images of 2-D gels for which information has been recorded in the lung protein database. This number represents a fraction of the 30 682 2-D gels produced by our group for different studies, which include studies of other cancer types encompassing head and neck, esophagus, liver, colon, pancreas, ovary, breast, prostate, brain, leukemias and childhood tumors. A list of protein gel images related to lung studies is shown in Table 2. While lung adenocarcinomas represent a major portion of the database, other lung tumor types including squamous cell carcinomas and small cell lung cancers are represented, as are control lung tissues. Other 2-D patterns were produced from studies of cell lines that have been manipulated by transfection or by treatment with specific agents, as well as patterns produced after different cell fractionation schemes. Substantial emphasis is currently being placed on the comprehensive profiling of lung cancer derived surface membrane proteins.

Table 2. A high-level categorization of lung protein 2-D images by sample type

Lung Sample Types	
Cell Lines	421
Cystic Fibrosis	44
Tumor	635
Normal	170
Plasma	61
Other	18
Total	1349

Mass spectrometry and/or *N*-terminal sequencing of protein spots from 2-D gels of lung tumor samples or cell lines have led to the identification of a large number of proteins expressed in lung cancer. Also, most identifications made for proteins from a sample type can often be confidently transferred to matching protein spots on master images from lung studies. Table 3 and Fig. 3 exhibit some of the progress we have made in identifying proteins in 2-D gels of lung samples.

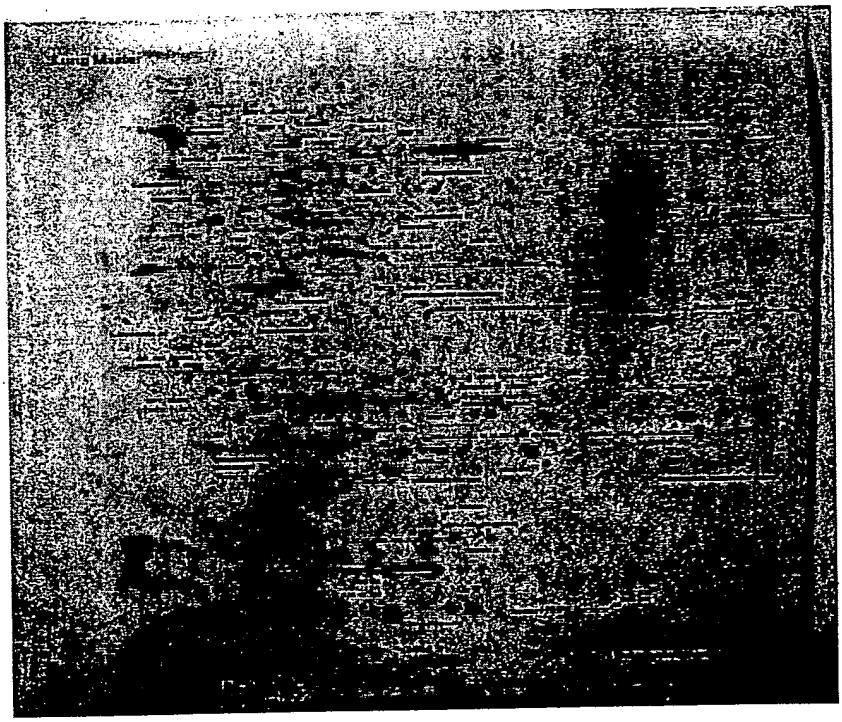


Figure 3. Small cell lung tumor master containing identified proteins.

Table 3. A list of identified proteins

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	*pI*	*M*,	Official gene
L95	(spot 1496L) possibly pacreatitis-associated protein	1496					
L M	14_3_3_sigma	577	398953	P31947	4.361	30.052	SFN
L M	14_3_3_ZetaDelta	615	112695	P29312	4.569	29.101	YWHAZ
L95	14–3–3η	1279	437363	AAA35483			YWHAH
DMS 79	6PF-2-K/FRU-2,6-P2ASE Liver isozymer	24	2507178	P16118			PFKFB1
	ADP-ribosylation factor 1	928	4502201	NP_001649	6.31	20.697	ARF1
	Albumin	319					ALB
	Albumin	800					ALB
L M	Albumin				5.957	70.244	ALB
L M	Aldehyde Dehydrogenase	207	4502031	NP_000680	6.811	56.966	ALDH1A1
L M	AldoKeto Reductase	543	3493209	AAC36469	7.812	32.379	AKR1B10
SKMES	Alkaline Phosphatase, Placental type 1 precursor	14	130737	P05187	5.86	57.954	ALPP
	Albumin	666					ALB
	Albumin	693					ALB
L M	α-Enolase	332	4503571	NP_001419	7.742	45.407	ENO1
L95	α-helical protein	268	8272482	AAFZ4221			HCR
L95	α-helix coiled-coil rod homologue	268	5360901	BAA82158			
L M	α Tublin	172	5174477	NP_006073	5.099	52.848	
L M	Amyloid B4A	802			4.796	17.194	
A549	Annexinl	460	113944	P04083	6.73	39.264	ANXA1
L M	Annexin V	522	4502107	NP_001145	4.83	33.326	ANXA5
L M	ApoAl	685			5.124	25.4	
L95	Apoprotein, pulmonary surfactant	1278	71967	LNHUPS			
L M	ARF1	795			6.33	18.909	ARF1
SKMES	β-Galactoside soluble lectin	96	227920	1713410A	5.34	14.584	LGALS1
LM	β-Actin	349	113270	P02570	5.29	41.7	ACTB
DMS 79	β-spectrin	61	29497	X59511			SPTB
LM	β Tubulin	229	4507729	NP_001060	4.75	49.8	TUBB
DMS 79	Calmodulin dependant phosphodiesterase	22	11995077	AB038211			
L M	Calreticulin	104	4757900	NP_004335	3.668	57.29	CALR
L M	Calreticulin32	469			3.442	48.772	
A549	CGI–46 protein	36	4929561	AAD34041	6.25	49.296	
A549	Chaperonin–like protein	149	4502643	NP_001753	7.034	60.547	CCT6A
L95	Clathrin light chain A	1338	4502899	NP_001824			CLTA
	Collagen, type XV, α 1	789					COL15A1
DMS 79	Complexin II	85	1362772	E57233			CPLX2
L M	Cellular retinoic acid-binding protein 2	856			5.415	11.858	CRABP2
L M	Cellular retinol-binding protein 1, CRPB1	855	4506451	NP_002890	4.667	10.297	RBP1
	Creatine kinase, brain	439	180570	AAC31758	5.34	42.618	CKB
L M	Cytochrome C bxydase VA	872			4.568	9.2	
A549	Cytokeratin 8	321	1673575	U76549			KRT8

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	pI	M_r	Official gene
A549	Cytokeratin 8	446	2506774	PO5787	5.52	53.674	KRT8
A549	Cytokeratin 8	439	2506774	P05787	5.52	53.674	KRT8
L M	Cytokeratin 15, keratin 15	289	4504915	NP_002266	4.153	49.261	KRT15
A549	Dihydrolipoamide dehydrogenase, mitochondrial precursor	759	118674	P09622			
L M	DJ1	811	6005749	NP_009193	6.44	21.015	DJ1
L M	DJ1_MER5	700			6.263	24.001	
DMS 79	dj475N16.1 (CTG4A)	57	6969163	CAB75301			
L M	DUTPhase	769			5.719	20.136	
L95	E2 ubiquitin-conjugating enzyme	1445	4885417	AB022435			HIP2
L M	EIF4d	718			5.104	22.961	
L M	EIF5A	839			4.599	10.957	
	Enhancer of rudimentary (Drosophila) homolog	902					ERH
	Enolase 2 (γ, neuronal)	295	119347	P09104	4.94	47.286	ENO2
L M	ENPL_HSP100	18			4.945	78.717	
A549	F1FO-type ATP synthase subunit d	1519	5453559	NP_0063475	5.21	18.491	ATP5JD
DMS 79	G1/S specific cyclin E1	31	3041657	P24864			CCNE1
L M	G3PD	540			7.457	31.772	
L M	γ-Actin	348	113278	P02571	5.146	42.315	ACTG1
L M	Glyoxalase1	650	417246	Q04760	4.833	25.572	GLO1
FMD 79	Granulocyte-macrophage colony-stimulating factor precursor	86	117561	PO4141			CSF2
L M	GRP75	87			5.9341	73.124	
L M	GRP78	79			5.187	68.109	
L M	GSTpi	690	726098	AAC13869	5.5	25.4	GSTP1
	Heat shock 27 kD protein 1	626	123571	PO4792	7.83	22.327	HSPB1
	Heat shock 27 kD protein 1	631	123571	PO4792	7.83	22.327	HSPB1
A549	Heterogeneous nuclear ribonucleoprotein H	457	5031753	NP_005511			HNRPH1
A549	HLA-B71 or HLB-B71 variant	818	511776	U11269	5.55	36.558	
L M	HSC70_HSP73	120			5.893	72.429	
L M	HSP90	46			5.276	76.096	
L95	HSPC089	1036	6841118	AAF28912			
L95	HSPC321	1547	6841292	AAF28999			
L95	HSPC321	1548	6841292	AAF28999			
A549	HuCha 60 SP 60	181	4504521	NP_002147	5.7	61	HSPD1
L95	Huntingtin associated protein	1595	1708113	P54255			HAP1
L95	Huntingtin associated protein	1548	1708113	P54255			HAP1
	Internexin neuronal intermediate filament protein, alpha	183	6225015	Q16352	5.48	54.908	INA
	Keratin 17	934	4557701	NP_00413	4.97	48.106	KRT17
DMS 79	KIAA1610 protein	26	10047295	AB046830			
L M	LamR	340			4.549	44.03	

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	p*I*	M_r	Official gene
	Lectin, galactoside-binding, soluble, 1 (galectin 1)	873	227920	1713410A	5.34	14.584	LGALS1
L M	Lipocortin	460	113944	PO4083	6.73	39.264	ANXA1
A549	L-Lactate Dehydrogenase H chain	906	126041	PO7195			LDHB
A549	L-lactate dehydrogenase H chain (LDH-B)	906	4557032	NP_002291			LDMB
L M	LaminB	924			5.787	69.625	
	Lymphocyte cytosolic protein 1 (L-plastin)	924	4504965	NP_002289	5.20	70.290	LCP1
L95	Macropain subunit zeta	1338	4506187	NP_002289			PSMAS
DMS 79	MHC class 1 histocompatability antigen protein	33	1236790	U06487			
DMS 79	Multicalaytic endopeptidase comples chain C2, long splice from	74	346314	JC1445	6.51	30.239	
L M	MyosinLightCahin3	815			4.11	15.172	
A549	Nm23, NDPKA	1456	127981	P15531	5.809	19.216	
	Non metastatic cells 1, protein (NM23A)	793	4557797	NP_000260	5.83	17.148	NME1
L M	Op 18, leukemia-associated phosphoprotein p18 (stahmin)	809	5031851	NP_005554	5.783	17.164	LAP18
L M	Op 18a	807	5031851	NP_005554	4.962	13.655	LAP18
L M	Op 18m	808	5031851	NP_005554	5.302	14.857	LAP18
L M	Phosphoglycerate MutB	639			7.083	27.227	
L M	Phospholipase C	248			5.7	56.5	
L M	PIMT	662			6.211	25.804	
L95	Pinch-2 protein	1695	9800509	AAF99328			
L95	Pinch-2 protein	1825	9800509	AAF99328			
L95	Possibly activin type II receptor precursor; DNA polymerase epsilon subunit B; or ITF-1 DNA binding protein	627					
L95	Possibly BTF2p44	1496					
A549	Possibly carbonci anhydrase III or UCH-L1; PGP 9.5	1242					
A549	Possibly δ-3,5 δ-2,4-Dienoly-CoA isomerase precusor	2138					
A427	Possibly G1 to S phase transition protein; serine-theonine phosphatease protein; or phosphatase 5 protein	321					
L95	Possibly GCF2 fusion protein or Bamacan homolog	320					
L95	Possibly glycosyltransferase	1519					
L95	Possibly HLA DQ	1271					

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	pI	M_r	Official gene
A549	Possibly hydroxyacylglutathione hydrolase or B-lymphocyte Antigen CD20	1080					
L95	Possibly microtubule-based motor protein	1438					
L95	Possibly putative novel protein similar to HPS	1427					
L95	Possibly Spi-B; unnamed protein product (AK001844); or protein kinase (γ15801)	1187					
L95	Possibly T-complex protein	630					
A549	Possibly U 1 small nuclear ribonuclear protein Á	1148					
L95	Possibly unnamed protein product (AK000369) or syntaxin	1064					
L95	Possibly unnamed protein product or Pro0282p protein	1351					
	procollagen-proline, 2-oxoglutarate 4-dioxgenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase; thyroid homone bindung protein p55)	110	2507460	PO7237	4.76	57.116	P4HB
	proliferating cell nuclear antigen	515	129697	P17070	4.4	37.5	PCNA
	Protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), β-isoform	104	5915686	P30154	4.84	66.202	PPP2R1B
L M	Protein H precursor	40			3.714	62.182	
L M	Protein kinase C inhibitor 1	882	4885413	NP_005331	7.714	11.521	H1NT
L95	Pulmonary surfactant apoprotein precusor	1278	190565	AAA36510			SFTPA1
L95	Pulmonary surfactant-associated protein	1278	131412	PO7714			SFTPA1
L M	R33729_1	848	3355455	AAC27824	7.508	13.163	
	Retinol-binding protein 1, cellular	855	4506451	NP_002890	4.99	15.850	RBP1
L M	RoSS_A_Antigen	69			3.215	47.903	
	S100 calcium-binding protein A11 (calgizzarin)	906					S100A11
	S100 calcium-binding protein A8 (calgranulin A)	910	115442	PO5109	6.51	10.834	S100A8
	S100 calcium-binding protein A9 (calgranulin B)	931	6094219	P50117	6.37	13.291	S100A9
DMS 79	Serine/threonine protein phosphatase 2A, 65kDa regulatory Subunit A, β isoform	14	5915686	P30154	4.84	66.202	PPP2R1B
	SET translocation (myeloid leukemia-associated)	376	1711383	Q01105	4.12	32.103	SET

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	pI	M_r	Official gene
	Small glutamine-rich tetraricopeptide repeat (TPR)-containing	476	8134666	O43765	4.81	34.063	SGT
	Stratifin	577	398953	P31947	4.68	27.774	SFN
L M	Superoxidedism CuZn	792	134611	P00441	5.6	17.3	SOD1
L M	Superoxide DismMN, superoxide dismutase 2, mitochondrial	737	134665	PO4179	7.887	20.78	SOD2
L M	TCP 1 β subunit	202			5.89	59.841	
L M	TCTP (translationally-controlled tumor protein 1)	680	4507669	NP_003286	4.688	25.143	TPT1
L M	Thioredoxin	896			4.689	9.207	
L M	Tplastin HSP 70	125			5.862	68.909	
L M	Transthyretin	842			5.693	14.714	
A549	Triosephosphate isomerase	672	136060	P00938	7.2	25.5	TPI1
L95	Tropomyosin, cytoskeletal type, tropomyosin 5	550	136096	P12324	4.5	31.9	
LM	Tropomyosin 4	548	13274400	AAK17926	4.377	32.733	TPM4
L95	Troponin T	866	408217	AAB27731			
L95	Troponin T	778	408217	AAB27731			
	Tublin, β polypeptide	229	4507729	NP_001060	4.78	49.907	TUBB
DMS 79	Tumor associated hydroquinone (NADH) oxidase tNOS	34	6644167	AF207881			
	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, epsilon polypeptide	576	1168198	P4266	4.63	29.174	YWHAE
	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	615	112695	P29312	4.73	26,645	YWHAZ
	Tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, theta polypeptide	579	112690	P27348	4.68	27.764	YWHAQ
A549	Ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase), UCH-L1; PGP 9.5, GST mu	656	136681	PO9936	5.283	27.745	UCHL1
L95	Unnamed protein product	1270	7023092	BAA91833			
A549	Urokinase plasminogen activator	842	487123	S39495	6.01	31.263	
L M	Vid1	293			4.712	47.485	
L M	Vid2	294			4.614	46.369	
L M	Vid4	337			4.464	45.322	
	Vimentin	294					VIM
A427	Vimentin	606	4507894	NM_003380			VIM
A549	Vimentin	505	418249	PO8670			VIM
A549	Vimentin	47	340234	M25246			VIM

In addition to 2-D gel analysis, most lung adenocarcinomas are examined at the genomic level using restriction landmark genome scanning, and by mutation analysis for a small number of genes. Transcriptomic analysis is done primarily using oligonucleotide microarrays, as part of our efforts to derive a molecular based classification of lung adenocarcinomas that is more predictive of clinical behavior for this group of tumors than current classification schemes. We also have similar molecular analyses of control lung tissue obtained from multiple sources including adjacent lung tissue from lung cancer patients as well as tissues obtained from non-cancer resected lung.

Only a fraction of the information in the 2-D patterns has been linked across all studies and analyses. The lung protein database contains the basic descriptive data of various samples analyzed, the images of the 2-D patterns that resulted from these samples, the quantitative spot data and information about which spots have been matched to each other, and conclusions or findings about spots. The database is intended to allow not only the retrieval of existing data, but also to mine new information and knowledge about protein expression in lung cells. Data mining activities consist, for example, of reviewing previous studies and finding out which 2-D gel patterns and protein spots are interesting for post-planned analysis and new discoveries. Such discoveries derive from: (1) identification of proteins that exhibit interesting expression profiles in 2-D patterns that have been regrouped from different experiments and studies; (2) expanded statistical analyses that cover protein expression patterns involving large numbers of experiments and images; (3) relating our data involving proteins to outside information; and (4) relating proteomic data to genomic data.

4 Use of the database for post planned analysis

4.1 Virtual matching

Interactive software packages are used to automatically detect and quantify spots and to match spots between different protein patterns, with visual editing to correct any errors in computer based matching. The spot match program has created indices that allow investigators to quickly navigate through many gels and easily compare spots on images from many different experiments and studies, discover proteins of interest, and access and view relevant data. Here the term "match" is used as a logical "transitive" relation, which means if spot A is matched to spot B and spot B is matched to spot C then the spots A and C are considered matched. The lung protein database contains data on proteins detected on various 2-D gels. Since all gels derived from whole cell or tissue lysates in the lung protein database are tied into a single hierarchy, protein identification data recorded for a spot is used to derive protein data for its matched spots using an advanced query capability of the database. This is known as "virtual matching" or "virtual protein identification", which allows investigators to access and view all matched images and the corresponding information from the lung protein database. With a click on a spot, one gets the result shown in Fig. 4. The virtual protein identification feature does not provide a 100% level of certainty of protein identification, but it makes possible the display of spots of interest. A combination of automated recognition and manual editing generally yields an accurate record of protein information in the database for previously unknown proteins. With this approach, the lung protein database will evolve and mature to include all correct data for further analysis and data mining.

4.2 Integrating protein spot data with MS data

As interest in proteomic analysis grows, a number of very large public databases are available to access protein data *via* the internet. Public databases offer a sophisticated text search and keyword search, which links any entered keyword to all protein information associated with that keyword, to ensure easy access to all relevant data. Protein identification using MALDI-MS relies on database searches and usually has three components: (1) peak detection which allows automatic determination of peptide masses; (2) search in protein sequence databases (SWISS-PROT and/or GenBank) for protein entries that match the masses; and (3) certainty calculation which determines the quality of the match for each protein in the list [4]. An example of such a software tool is the PepFrag for searching protein and DNA sequence databases that can use different types of mass spectrometric information [5]. Fenyo [6] described methods and software tools in proteomics for identifying and characterizing proteins, which emphasizes MS combined with database searching. Proteolytic peptide mapping and genome database searching provide an automated means for identifying proteins, and the certainty of the results is computed by the number of masses matched for each protein [7]. Another useful tool is FindMod (http://www.expasy.ch/sprot/findmod.html) for the systematic characterisation of proteins using mass spectrometry [8].

We have created MS data forms that contain information used in mass spectrometry queries, summary information (Rank, MOWSE score, % Masses Matched, MW, p*I*, Species, Accession #, Protein Name) and additional information (Summary ID, Submitted Mass, Matched Mass, Delta

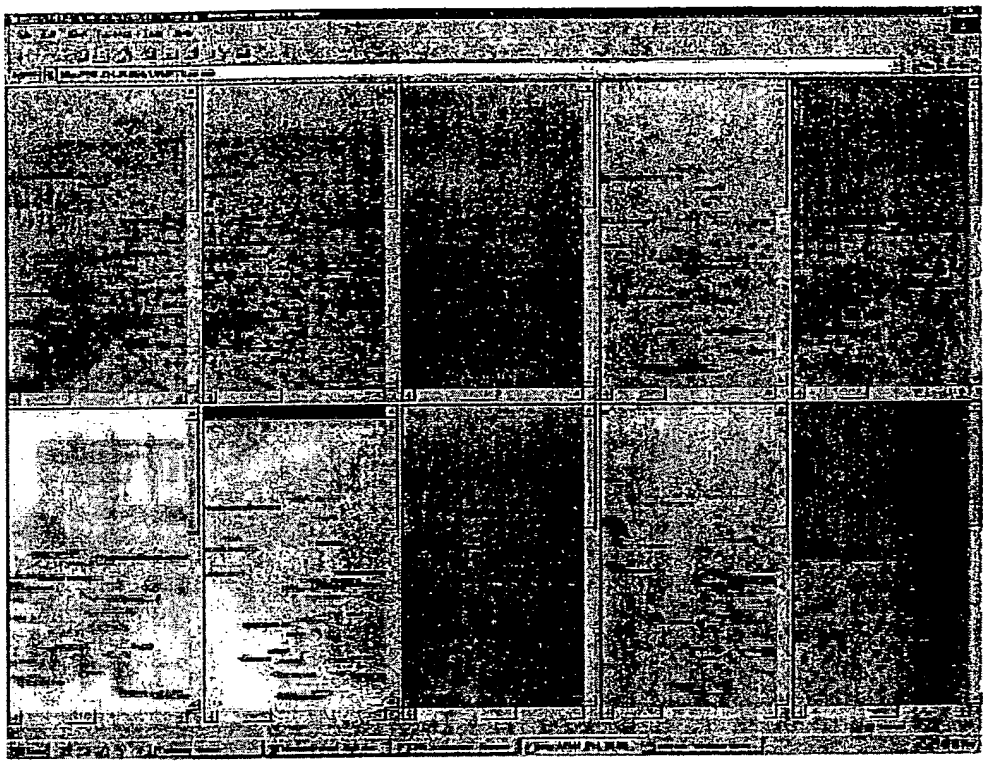


Figure 4. Virtual protein identification by clicking a spot.

PPM, Start, End, Peptide Seq, Modifications, Unmatched Masses). An example of the MS data form is shown in Fig. 5. Integrating the lung protein database with MS data provides a record of protein identification and high level of integration with other public databases, although substantial effort is required for data collection. We are currently evaluating an automated or semi-automated method of pulling these data when new information, which is relevant to our objectives, is available.

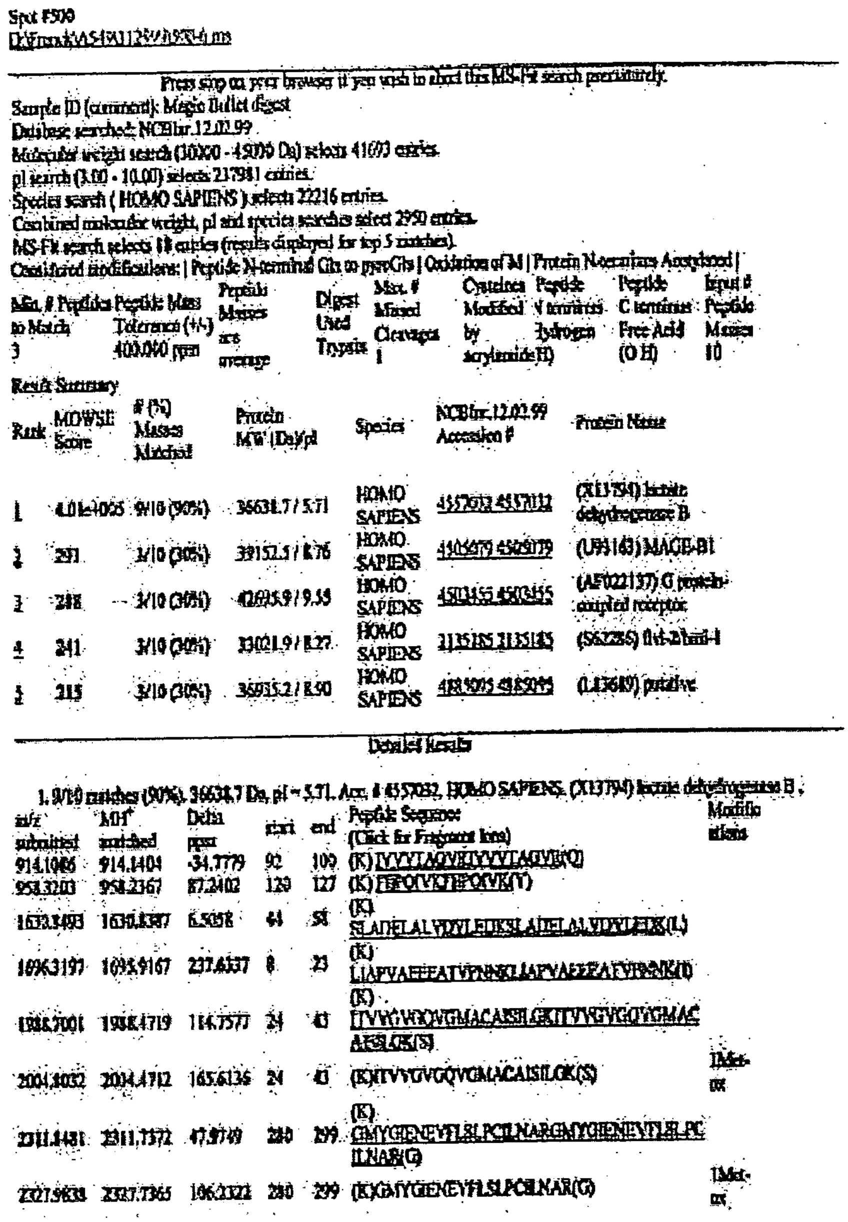
4.3 Integrating protein data with microarray data

As technology evolves, new computer aids and methods are introduced for genomic analysis as well as proteomic analysis. With respect to DNA microarray platforms, a current goal is to construct lung specific cDNA microarrays for lung cancer investigations. In the meantime RNA expression data for lung cancer is being collected using an Affymetrix oligonucleotide based system. This system automates the identification and quantification of microarray spots. Data files contain integrated intensities for each spot and ratios showing fold changes *per* spot. The use of oligonucleotide based microarrays for RNA analysis in lung cancer by our group has resulted in a massive amount of data. Integration of protein information in the lung protein database with microarray data allows us to extend data analysis capability to encompass RNA and protein data for a subset of genes.

5 Some findings derived from the lung cancer protein database

5.1 Unique proteomic pattern of small cell lung cancer

A major goal of our proteomic and genomic studies of lung cancer is to derive novel classification schemes that have utility in making a diagnosis, predicting outcome and in making therapeutic decisions. An important first step in this direction is to determine the ability of proteomic profiling to distinguish between known types of lung cancer. Specific protein differences between different types of cancer have been identified by other groups. In a recent study of breast, ovary and lung tumors, 20 differentially expressed proteins were identified [9] and in a prior study, 16 polypeptides were found to be associated with different histopathological features of lung cancer [10, 11]. In a study of 25 adenocarcinomas of the lung, 12 small cell



Spot #500
[illegible]

Press stop on your browser if you wish to abort this MS-Fit search prematurely.
Sample ID (comment): Magic Bullet digest
Database searched: NCBInr.12.02.99
Molecular weight search (10000 - 45000 Da) selects 41603 entries.
pI search (3.00 - 10.00) selects 217981 entries.
Species search (HOMO SAPIENS) selects 22216 entries.
Combined molecular weight, pI and species searches select 2990 entries.
MS-Fit search selects 18 entries (results displayed for top 5 matches).
Considered modifications: | Peptide N-terminal Gln to pyroGlu | Oxidation of M | Protein N-terminus Acetylated |

Min. # Peptides to Match	Peptide Mass Tolerance (+/-)	Peptide Masses are	Digest Used	Max. # Missed Cleavages	Cysteines Modified by	Peptide N terminus	Peptide C terminus	Input # Peptide Masses
3	400.000 ppm	average	Trypsin	1	acrylamide	Hydrogen (H)	Free Acid (O H)	10

Result Summary

Rank	MOWSE Score	# (%) Masses Matched	Protein MW (Da)/pI	Species	NCBInr.12.02.99 Accession #	Protein Name
1	4.01e+005	9/10 (90%)	36631.7 / 5.71	HOMO SAPIENS	4557032 4557032	(X13794) lactate dehydrogenase B
2	251	3/10 (30%)	39152.5 / 8.76	HOMO SAPIENS	4505079 4505079	(U93163) MAGE-B1
3	248	3/10 (30%)	42035.9 / 9.55	HOMO SAPIENS	4503455 4503455	(AF022137) G protein-coupled receptor
4	241	3/10 (30%)	33021.9 / 8.27	HOMO SAPIENS	3135185 3135185	(S62286) [illegible]
5	215	3/10 (30%)	36935.2 / 8.90	HOMO SAPIENS	4885095 4885095	(L13689) putative

Detailed Results

1. 9/10 matches (90%). 36638.7 Da, pI = 5.71. Acc. # 4557032. HOMO SAPIENS. (X13794) lactate dehydrogenase B .

m/z submitted	MH+ matched	Delta ppm	start	end	Peptide Sequence (Click for Fragment Ions)	Modifications
914.1086	914.1404	-34.7779	92	109	(K) [illegible] (Q)	
958.3200	958.2367	87.2402	120	127	(K) [illegible] (T)	
1633.3493	1633.3387	6.5058	44	58	(K) SLADELALVDVLEDKSLADELALVDVLEDK (L)	
1096.3197	1095.9167	237.6337	8	23	(K) LIAPVAEEEATVPNNKLIAPVAEEEATVPNNK (I)	
1988.7001	1988.4719	114.7577	24	43	(K) ITVVGVGQVGMACAISILGKITVVGVGQVGMACAISILGK (S)	
2004.4032	2004.4712	165.6136	24	43	(K) ITVVGVGQVGMACAISILGK (S)	1Met-ox
2311.4481	2311.7372	47.9749	280	299	(K) GMYGIENEVFLSLPCILNARGMYGIENEVFLSLPCILNAR (G)	
2327.9838	2327.7365	106.2322	280	299	(K) GMYGIENEVFLSLPCILNAR (G)	1Met-ox

Figure 5. MS data form

lung cancers, and 16 squamous cell tumors, by our group (manuscript submitted) an initial analysis of protein 2-D patterns uncovered a group of 52 protein spots that differed in average integrated intensity between the three groups. Performing simple two-sample *t*-tests gave *p* values of less than 0.05 for the 52 spots for at least one of the pairs of groups. Most of the spots differed between small cell and the remaining two diagnostic groups, with 47 spots differing significantly between small cell and adenocarcinoma groups and 44 between small cell and squamous ($p<0.05$). Between the adenocarcinoma and squamous groups 12 spots with difference of this significance were found. Summary data for some of the spots is presented in Table 4. The first two principal components of the data are graphed in Figure 6, and show that as a group the spots distinguish small cell tumors from the other two tumor types fairly easily.

We have identified 39 of this set of 52 spots by either *N*-terminal sequencing and/or MS of spot digests. Small cell lung cancers were characterized by higher average amounts for some proteins associated with cell prolifera-

Table 4. 39 identified protein spots found to differ between small cell, adenocarcinoma, and squamous tumors of the lung ($n = 12, 25, 16$). In the *t*-test columns are p values from the two–sided two–sample *t*-test comparing each pair of groups

Spot #	Unigene description	Offical gene symbol	Mean adeno-car-cinoma	Mean squa-mous	Mean small cell	*t*-test adenocar-cinoma *vs* small cell	*t*-test small cell *vs* squa-mous	*t*-test adenocar-cinoma *vs* squamous
294	vimentin	VIM	1.36	1.16	0.53	0.010	0.016	0.509
319	albumin	ALB	2.13	1.67	0.73	0.001	0.005	0.231
666	albumin	ALB	0.72	0.59	0.20	0.002	0.030	0.461
800	albumin	ALB	2.34	1.80	0.63	0.010	0.034	0.383
873	lectin, galactoside-binding, soluble, 1 (galectin 1)	LGALS1	1.95	1.69	0.83	0.000	0.002	0.310
928	ADP-ribosylation factor 1	ARF1	0.22	0.19	0.06	0.012	0.046	0.607
522	annexin A5	ANXA5	0.46	0.26	0.39	0.429	0.202	0.012
515	proliferating cell nuclear antigen	PCNA	0.15	0.18	0.36	0.002	0.011	0.464
577	stratifin	SFN	0.78	1.39	0.41	0.129	0.002	0.029
626	heat shock 27 kD protein1	HSPB1	0.87	1.18	0.30	0.000	0.002	0.128
631	heat shock 27 kD protein1	HSPB1	1.04	1.35	0.46	0.003	0.017	0.277
793	non-metastatic cells 1, protein (NM23A)	NME1	0.36	0.43	0.59	0.003	0.033	0.253
807	leukemia-associated phosphoprotein p18 (stathmin)	LAP 18	0.03	0.05	0.92	0.000	0.000	0.351
809	leukemia-associated phosphoprotein p18 (stathmin)	LAP18	0.55	0.50	3.88	0.000	0.000	0.732
931	S100 calcium-binding protein A9 (calgranulin B)	S100A9	0.95	1.18	0.24	0.026	0.001	0.447
104	protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform	PPP2R1B	0.17	0.13	0.65	0.000	0.001	0.188
110	procollagen-proline, 2oxoglutarate 4-dioxy-genase (proline 4-hydroxylase) beta polypeptide (protein disulfide isomerase; thyroid hormone binding protein p55)	P4HB	0.10	0.10	0.30	0.014	0.049	0.906
183	internexin neuronal intermediate filament protein, alpha	INA	0.04	0.04	0.16	0.000	0.000	0.751
229	tubulin, beta polypeptide	TUBB	0.14	0.27	0.83	0.000	0.000	0.028
289	keratin 15	KRT15	0.36	0.29	0.65	0.028	0.009	0.343
295	enolase 2, (gamma, neuronal)	ENO2	0.10	0.23	0.39	0.000	0.065	0.026
376	SET translocation (myeloid leukemia-associated)	SET	0.25	0.17	0.71	0.000	0.000	0.031
439	creatine kinase, brain	CKB	0.11	0.05	0.16	0.033	0.000	0.004
460	annexin A1	ANXA1	0.43	0.42	0.59	0.014	0.026	0.691
476	small glutamine-rich tetratricopeptide repeat (TPR)-containing	SGT	0.16	0.19	0.33	0.000	0.000	0.241
576	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase acctivation protein, epsilon polypeptide	YWHAE	0.40	0.38	0.82	0.000	0.001	0.697
579	tyrosine 3-monooxygenase/trypthophan 5-monooxgenase activation protein, theta polypeptide	YWHAQ	0.52	0.55	0.91	0.000	0.006	0.703
615	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	YWHAZ	0.93	1.09	1.79	0.000	0.003	0.336

Table 4. Continued

Spot #	Unigene description	Offical gene symbol	Mean adeno-car-cinoma	Mean squa-mous	Mean small cell	*t*-test adenocar-cinoma *vs* small cell	*t*-test small cell *vs* squa-mous	*t*-test adenocar-cinoma *vs* squamous
656	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	UCHL1	0.17	0.32	0.85	0.000	0.005	0.153
855	retinol-binding protein 1, cellular	RBP1	0.42	0.41	0.77	0.006	0.014	0.961
856	cellular retinoic acid-binding protein 2	CRABP2	0.25	0.38	0.63	0.000	0.017	0.037
902	enhancer of rudimentary (Drosophila) homolog	ERH	0.38	0.35	0.76	0.000	0.000	0.455
910	S100 calcium-binding protein A8 (calgarnulin A)	S100A8	1.46	1.43	0.35	0.040	0.001	0.950
934	keratin 17	KRT17	0.16	0.30	0.15	0.768	0.073	0.013
693	allbumin	ALB	2.63	1.98	0.92	0.000	0.008	0.138
737	superoxide dismutase 2, mitochondrial	SOD2	1.17	1.22	0.54	0.013	0.001	0.836
789	collagen, type XV, alpha 1	COL15A1	0.57	0.50	0.26	0.031	0.186	0.658
906	S100 calcium-binding protein A11 (calgizzarin)	S100A11	2.95	2.62	0.53	0.000	0.000	0.506
924	lymphocyte cytosolic protein 1 (L-plastin)	LCP1	0.18	0.13	0.05	0.000	0.004	0.034

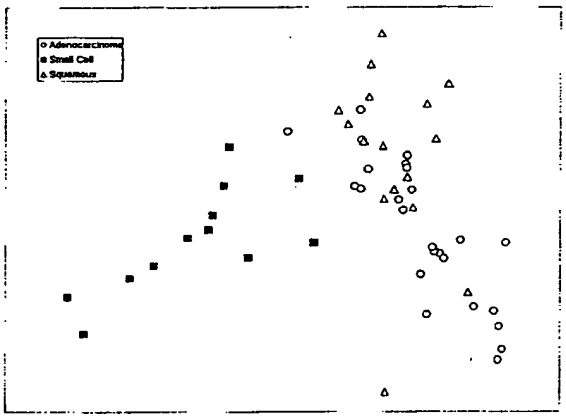


Figure 6. First two principle components for 52 protein spots distinguishing between lung tumor types. Small cell lung cancer samples are shown as squares, adenocarcinomas as circles and squamous lung tumors as triangles.

tion such as proliferating cell nuclear antigen (PCNA) and oncoprotein 18 (Op18) [12–15], particularly the once-phosphorylated form of Op18, as well as protein products of the UCHL1, RBP1, CRABP2, KRT15, and TUBB genes among others. Squamous cell and adenocarcinoma samples had greater amounts of the S100 proteins S100A8, S100A9, and S100A11, as well as larger average amounts of both the unphosphorylated and phosphorylated 27 kD heat shock protein (HSPB1). These two groups also had larger amounts of several protein spots detected on these gels that did not occur in similar gels made from cell lines and were thought to be cleavage products from proteins present in cells or plasma surrounding the tumor cells (*e.g.* cleaved albumin). The number of protein spots that differed between lung adenocarcinomas and squamous tumors were fewer than the number of proteins that distinguished between small cell lung cancer and the other two lung cancer types. ENO2 was smallest in the adenocarcinoma group, while ANXA5 and CKB were lowest and KRT17 and SFN highest in the squamous carcinoma samples. Several interesting spots found in the study remain to be definitively identified.

5.2 Correlations between RNA and protein expression

The availability of mRNA expression data from microarrays or Affymetrix chips for the same samples for which we have protein 2-D gel data permits several additional types of questions to be asked. We have thus far entertained only simple models of protein/mRNA relationships that ask which mRNA levels are most correlated with protein spot sizes. Figure 7 depicts such a correlation matrix using colors rather than numerical data, since this makes it easier to visualize the relationships. In cases for which the identity of the protein spot is known such investigations can answer the question of how well mRNA levels for a protein predict that protein's abundance. In cases of protein spots that have not yet been identified, or iden-

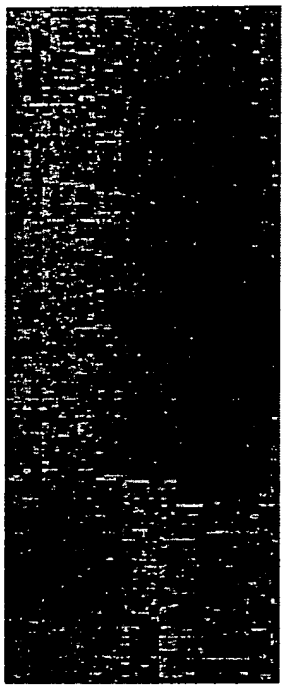


Figure 7. Correlation matrix of 30 protein spots (columns) with mRNA levels as measured by 200 probe-sets on Affymetrix HuFL chips. The correlation coefficients are depicted with colors, bright red being near-perfect correlation ($r = 1$) and bright green anticorrelation ($r = -1$). The figure was made using the TreeView software (rana.lbl.gov/EisenSoftware.htm).

tified without high confidence, such correlations can lead to or confirm hypothetical spot identifications. More generally one can search for larger groups of proteins and mRNA whose abundances are controlled by some common mechanism.

5.3 Identification of novel lung cancer markers

We have utilized a proteomic approach to identify proteins that commonly induce an antibody response in lung cancer. Such identified proteins or their corresponding autoantibodies likely have substantial utility for cancer diagnosis. There is also evidence that autoantibodies may be present prior to clinical diagnosis and therefore detection of autoantibodies or of circulating antigens may have utility for screening and early diagnosis of cancer. We have identified a battery of proteins that induce autoantibodies that are specific for different types of cancer. We have identified a panel of autoantibodies that are detectable in serum of lung cancer patients at the time of diagnosis. The availability of a database of protein expression in lung cancer has facilitated the identification of proteins that induce autoantibodies in addition to providing valuable information regarding the expression pattern of such antigens in different tumor types and cell lines. One such antigen we have identified in lung cancer is protein PGP 9.5 (Fig. 8) (Brichory *et al*, manuscript submitted) [16]. PGP 9.5 was identified as a protein in lung cancer that induces autoantibodies as part of a study in which sera from 64 newly diagnosed patients with lung cancer, from 99 patients with other types of cancer and from 71 noncancer controls were analyzed for antibody-based reactivity against lung adenocarcinoma proteins resolved by 2-D PAGE. Gels containing separated proteins were blotted and subsequently hybridized with individual sera from patients or controls. Unlike controls, autoantibodies against a protein identified by MS as protein gene product 9.5 (PGP 9.5) were detected in sera in 9 out of 64 patients with lung cancer.

Circulating PGP 9.5 antigen was detected in sera from two additional patients with lung cancer, without detectable PGP 9.5 autoantibodies. PGP 9.5 is a neurospecific polypeptide previously proposed as a marker for nonsmall cell lung cancer, based on its expression in tumor tissue. Using A549 lung adenocarcinoma cell line, we have demonstrated that PGP 9.5 was present at the cell surface, as well as secreted. Thus, the findings of PGP 9.5 antigen and/or antibodies in serum of patients with lung cancer suggest that PGP 9.5 may have utility in lung cancer screening and diagnosis, as part of a panel of such proteins or their corresponding antibodies, which we have identified.

6 Web pages

The relational database for storage of sample, image, protein information and other related data is being constructed in a stepwise fashion. The construction of a comprehensive database to collect all pertinent information is rather challenging and necessitates substantial resources. Similar effort in this area includes WebGel that is a web based gel database analysis system that contains previously quantified gel data generated from a stand-alone quantitative gel analysis system [16]. Public WebGel demonstration databases currently available can be found in the web site (http://www-lecb.ncifcrf.gov/webgel WebGel database). The task of web based retrieval of data from the protein database is rather complex as there are different kinds of data that may need to be retrieved. The microarray data could be stored in the database instead of Excel files, and the Access 2000 database that the MS team utilizes could be transferred to the database. Tables are being built to eliminate any handwritten collection of data. Developing a database is

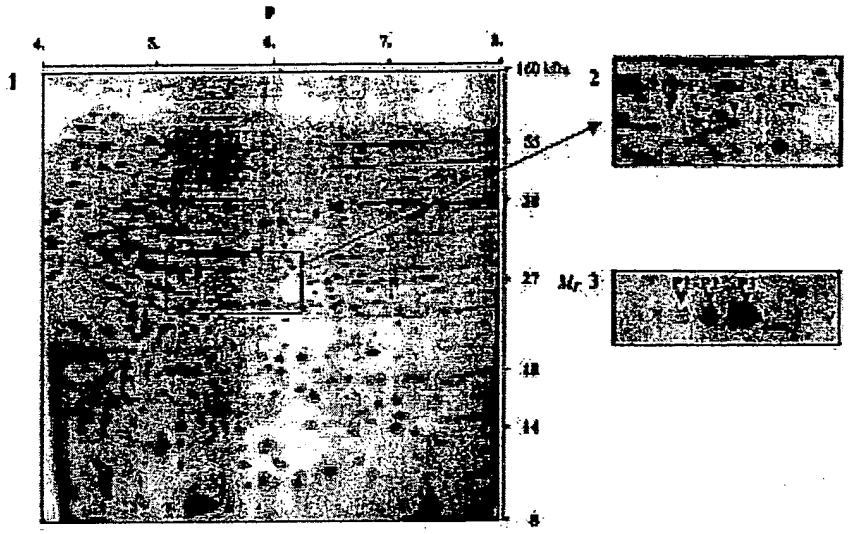


Figure 8. 2-D PAGE and Western blot analysis of A549 lung adenocarcinoma cell proteins. Panel 1 shows A549 2-D protein pattern after silver staining. The boxed area is shown in panel 2, in which arrows point to the location of PGP 9.5 forms (spots P1 to P3) recognized by sera from patients with lung cancer and the position of the form P4 recognized by a polyclonal rabbit anti-PGP 9.5 antiserum, which also recognizes P1-P3. Panel 2 shows close-ups of western blots hybridized with two different sera from patients with lung adenocarcinoma that showed reactivity against PGP 9.5 proteins.

hard because of complex and very large amount of unstructured data generated. There are conflicting pressures between "using what we've got already" and constructing something better. Sometimes there is a natural break in the data, such as when a shift is made from one platform type to another. Then one could "pile up" old data and organize it neatly. On the other hand, when new technologies are introduced, they require new ways of storing the data. The lung protein database is continuously evolving to enhance the relational schema to be more flexible and comprehensive and to make data processing more robust and automatic.

The lung protein database is a backbone to record proteome data for many different studies and to mine the existing data for new discoveries. The new generation LIPS provides investigators web-enabled interfaces to the laboratory databases and 2-D images with internet access. There is certainly a need for sharing information in the database on a global basis. We have used internet and WWW technologies to provide a distributed process with easy-to-use front-end user interface. Figure 9 shows a top level view of a web-based process for performing our studies from a data processing perspective. Some of our web pages were developed in Visual InterDev and ASP development environment on Microsoft and some were developed in Oracle 8i and WebDB web application environment on Solaris. As an example, the MS data web page is shown in Fig. 10. Detailed "how-to" documentation is provided as on-line help for recently extended capabilities of LIPS.

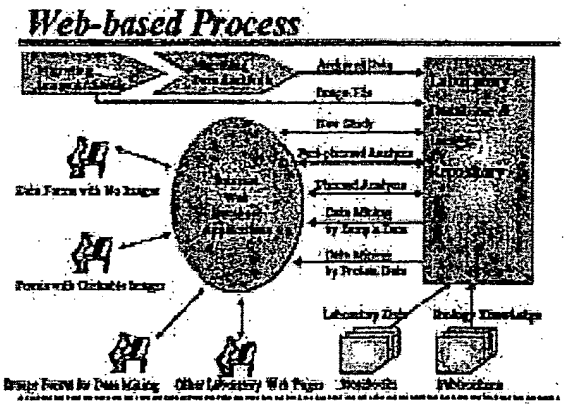


Figure 9. Web-based process of using lung protein database.

7 Conclusion

The value of the database we have constructed depends to a large measure on its content, the quality of data and the ease with which data can be retrieved and analyzed. While the amount of data generated is already quite sizeable, it is likely that the database will continue to undergo substantial expansion. Proteins are being identified at a rapid pace, thus enhancing our ability to link protein expression data with RNA based expression data for corresponding genes. As such, the database will play an important role in achieving our objective of developing novel classification schemes for lung cancer and the identification of novel markers for early diagnosis. The

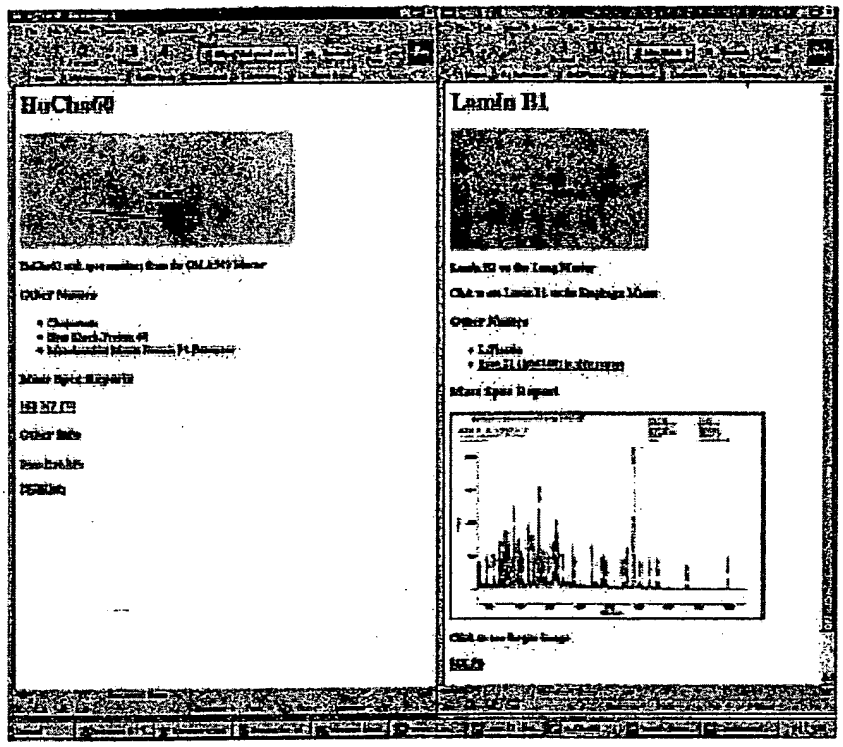


Figure 10. MS data web page.

database will also serve as a useful resource for other investigations of lung biology and of diseases other than lung cancer.

Received May 20, 2001

8 References

[1] Lemkin, P. F., *Electrophoresis* 1997, *18*, 2759–2773.

[2] Ali, I., Chan, Y., Kuick, R., Teichroew, D., Hanash, S. M., *Electrophoresis* 1991, *12*, 747–761.

[3] Oh, J. M. C., Hanash, S. M., Teichroew, D., *Electrophoresis* 1999, *20*, 766–774.

[4] Gras, R., Muller, M., Gasteiger, E., Gay, S., *et al.*, *Electrophoresis* 1999, *20*, 3535–3550.

[5] Fenyo, D., Qin, J., Chait, B. T., *Electrophoresis* 1998, *19* 998–1005.

[6] Fenyo, D., *Curr. Opin. Biotechnol* 2000, *11*, 391–395.

[7] Eriksson, J., Chait, B. T., Fenyo, D., *Anal. Chem.* 2000, *72*, 999–1005.

[8] Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., *et al.*, *J. Mol. Biol.* 1999, *289*, 645–657.

[9] Bergman, A. C., Benjamin, T., Alaiya, A., Waltham, M., *et al.*, *Electrophoresis* 2000, *21*, 679–686.

[10] Hirano, T., Franzen, B., Uryu, K., Okuzawa, K., *et al.*, *Br. J. Cancer* 1995, *72*, 840–848.

[11] Schmid, H. R., Schmitter, D., Blum, P., Miller, M., Vonderschmitt, D., *Electrophoresis* 1995, *16*, 1961–1968.

[12] Wang, Y. K., Liao, P.-C., Allison, J., Gage, D. A., *et al.*, *J. Biol. Chem.* 1993, *268*, 14269–14277.

[13] Melhem, R. F., Zhu, X. X., Hailat, N., Strahler, J., Hanash, S. M., *J. Biol. Chem.* 1991, *266*, 17747–17753.

[14] Zhu, X. X., Kozarsky, K., Strahler, J. R., Eckerskorn, C., *et al.*, *J. Biol. Chem.* 1989, *264*, 14556–14560.

[15] Hanash, S. M., Strahler, J. R., Kuick, R., Chu, E. H. Y., Nichols, D., *J. Biol. Chem.* 1988, *263*, 12813–12815.

[16] Lemkin, P. F., Myrick, J. M., Lakshmanan, Y., Shue, M. J., *et al.*, *Electrophoresis*. 1999, *20*, 3492–3507.

Translational regulation of human p53 gene expression

Loning Fu, Mark D.Minden and Sam Benchimol

The Ontario Cancer Institute/Princess Margaret Hospital and Department of Medical Biophysics, University of Toronto, 610 University Avenue, Toronto, Ontario, Canada M5G 2M9

In blast cells obtained from patients with acute myelogenous leukemia, p53 mRNA was present in all the samples examined while the expression of p53 protein was variable from patient to patient. Mutations in the p53 gene are infrequent in this disease and, hence, variable protein expression in the majority of the samples cannot be accounted for by mutation. In this study, we examined the regulation of p53 gene expression in human leukemic blasts and characterized the p53 transcripts in these cells. We found control both at the level of RNA abundance and at the level of translation. Four experiments point towards translational control of human p53 gene expression. First, there is no correlation between the level of p53 mRNA and the level of p53 protein expression in blast cells. Second, in two cell lines with similar levels of p53 protein expression but with different levels of p53 mRNA, we find that there is preferential association of p53 mRNA with large polysomes in the cells with less p53 RNA. Third, translation of synthetic human p53 transcripts in cell-free extracts is inhibited by the p53 3'UTR. Fourth, the p53 3'UTR, when present *in cis*, can repress translation of a heterologous transcript. These observations raise the possibility that human p53 mRNA translation may be regulated *in vivo* by RNA binding factors acting on the p53 3'UTR.

Keywords: acute myelogenous leukemia/p53/translational control

Introduction

Human acute myelogenous leukemia (AML) is a clonal disease arising in a very early hematopoietic progenitor cell following multiple carcinogenic events (Wiggans *et al.*, 1978; Fialkow *et al.*, 1987). Mutation of the p53 tumor suppressor gene occurs infrequently in the blast cells of AML patients (Fenaux *et al.*, 1991, 1992; Slingerland *et al.*, 1991; Sugimoto *et al.*, 1991, 1993; Zhang *et al.*, 1992; Trecca *et al.*, 1994; Wattel *et al.*, 1994; Lai *et al.*, 1995). p53 mutations have been detected in ~10% of all AML patients, mostly in patients with 17p monosomy who had lost the normal remaining p53 allele (Lai *et al.*, 1995). These studies demonstrate that p53 mutations are not required for the development of AML. Mutations that do arise, however, are generally recessive in nature, indicating a strong selective pressure to eliminate completely wild-type p53 protein function.

The scarcity of p53 gene mutations in AML is not unique to this disease. For example, p53 gene mutations are rare in neuroblastoma, testicular tumors and HPV-positive cervical cancer. While the p53 gene is most commonly inactivated through mutation in human tumors, p53 protein function can also be disrupted through non-genetic mechanisms including protein–protein interactions (Scheffner *et al.*, 1990; Momand *et al.*, 1992; Oliner *et al.*, 1992; Ueda *et al.*, 1995), protein conformational change (Milner, 1991; Ullrich *et al.*, 1992) and nuclear exclusion (Moll *et al.*, 1992, 1995). Indeed, two groups have suggested that inactivation of wild-type p53 protein in AML occurs through a mechanism involving conformational change of the protein (Zhu *et al.*, 1993; Zhang *et al.*, 1992).

The level of p53 protein expression in primary blast cells obtained from AML patients varies from patient to patient. In previous studies from this laboratory p53 protein expression was detected in only 45% (34 of 75) blast samples examined by metabolic labelling with [^{35}S]methionine and immunoprecipitation (Smith *et al.*, 1986; Benchimol *et al.*, 1989; Slingerland *et al.*, 1991). Zhang *et al.* (1992) detected p53 protein expression in blast samples from 75% (37 of 49) AML patients. Several reasons may explain the absence or very low level of p53 protein expression in certain blast samples. These include low levels of p53 mRNA, inhibition of p53 mRNA translation and extremely rapid turnover of newly synthesized p53 protein. In this study, we have examined the regulation of p53 gene expression in human AML blasts and find control both at the level of RNA abundance and at the level of translation. Translational regulation is supported by experiments in which we demonstrate that the p53 3' untranslated region (3'UTR) can repress translation of p53 RNA and of heterologous transcripts in cell-free extracts.

Results

Expression of p53 protein in human AML

Leukemic blast cells from AML patients and three human acute leukemia cell lines OCI-M2, OCI/AML-3 and OCI/AML-4 were characterized for p53 protein expression by metabolic labelling and immunoprecipitation. OCI-M2 is a human erythroleukemia cell line (Papayannopoulou *et al.*, 1988) previously shown to contain a missense mutation in the p53 coding region at codon 274 and to have lost the homologous wild-type p53 allele (Slingerland *et al.*, 1991). OCI/AML-3 and OCI/AML-4 cell lines were derived from the primary blasts of two AML patients (Wang *et al.*, 1989). The full-length p53 transcripts in these cells were amplified by RT–PCR, and the products directly sequenced. We found that the p53 transcripts in both cell lines were wild-type throughout their coding

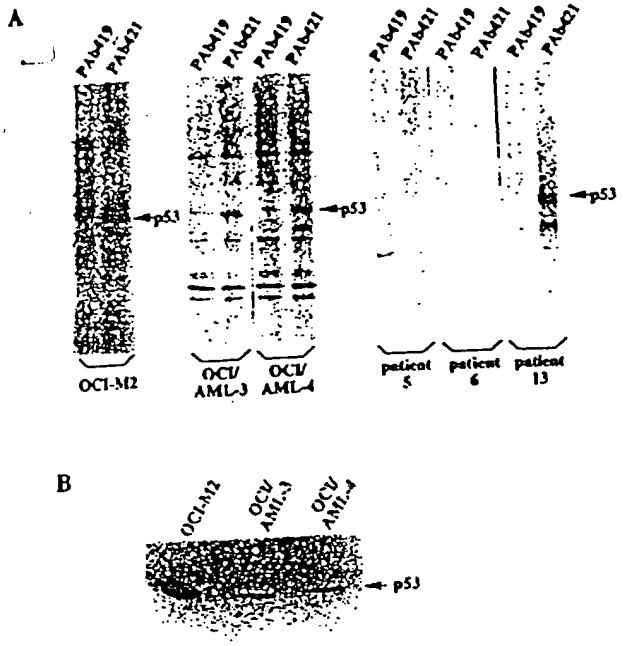


Fig. 1. Expression of p53 protein in human leukemia cells. (A) Cell lines OCI-M2, OCI/AML-3 and OCI/AML-4, and blast cells from AML patients were metabolically labelled with [^{35}S]methionine for 15 min at 37°C. Cell extracts were prepared and portions representing equal amounts of trichloroacetic acid-insoluble radioactivity (10^7 c.p.m.) were immunoprecipitated with the control monoclonal antibody (PAb419) or with monoclonal antibodies against p53 (PAb421). (B) Detection of p53 protein in 5×10^6 cells by Western immunoblotting and ECL using PAb1801 monoclonal antibodies.

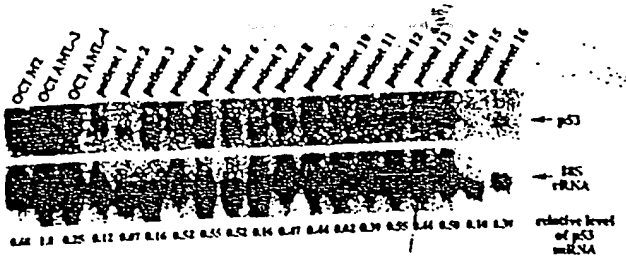


Fig. 2. Northern blot analysis of p53 mRNA in human AML cells. 20 µg of total RNA isolated from cell lines or patient blast samples was separated on a 1% agarose gel containing 6% formaldehyde, transferred to nitrocellulose and hybridized with ^{32}P-labelled human p53 cDNA. After autoradiography, the probe was removed and the filters were hybridized with a probe specific for 18S ribosomal RNA. The relative abundance of p53 mRNA was determined by phosphorimage analysis after normalizing to the value of 18S ribosomal RNA in each sample.

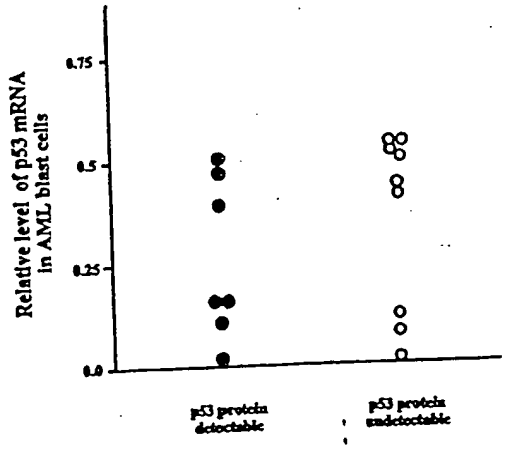


Fig. 3. Relative abundance of p53 mRNA in cells that do or do not express detectable p53 protein. p53 protein synthesis was assessed in 16 AML blast samples by metabolic labelling with [^{35}S]methionine for 15 min and immunoprecipitation. p53 protein synthesis was detected in seven of these samples. p53 mRNA levels were determined by Northern blot analysis as described in the legend to Figure 2.

regions as well as through their 5′- and 3′UTRs. The only difference detected in the p53 transcripts expressed in OCI/AML-3 and OCI/AML-4 cell lines was the recognized polymorphism at codon 72 (Matlashewski *et al.*, 1987) resulting in an arginine residue in OCI/AML-3 and a proline residue in OCI/AML-4 at position 72.

The level of protein expression measured by metabolic labelling and immunoprecipitation is dependent primarily on the rate of protein synthesis, the rate of protein degradation and the amount of mRNA available for translation. To minimize the contribution of protein half-life on the detection of p53 protein synthesis during the metabolic labelling assay, cells were exposed to a short 15 min pulse of [^{35}S]methionine at 37°C followed by immediate lysis on ice in the presence of protease inhibitors. Radiolabelled cell extracts prepared in this way were then subjected to immunoprecipitation with p53-specific antibodies. p53 protein with a half-life much less than 15 min, however, might remain undetectable by this assay. p53 protein synthesis was detected in OCI/AML-3, OCI/AML-4 and in OCI-M2 (Figure 1A) as well as in seven of 16 blast samples tested; three representative examples are shown in Figure 1A.

The steady-state level of p53 protein in the three cell lines was determined by Western blot analysis using PAb1801. Densitometric scanning of the blot shown in Figure 1B revealed that the amount of p53 protein in OCI/AML-3 and OCI/AML-4 was similar and ~10-fold lower than in OCI-M2. The high level of p53 protein in OCI-M2 was expected since mutant p53 polypeptides usually have much longer half-lives than wild-type p53 proteins and as a result mutant p53 polypeptides accumulate intracellularly.

Expression of p53 mRNA in human AML

To determine whether the differences in p53 protein expression in leukemic blasts reflected differences in the abundance of p53 mRNA, RNA was isolated from AML blast samples and cell lines, and subjected to Northern blot analysis. The relative abundance of p53 mRNA in cells was estimated by phosphorimage analysis after normalizing to the value of 18S ribosomal RNA in each sample. The results are shown in Figure 2 and indicate that the 16 AML blast samples examined synthesized a single species of full-length p53 mRNA ~2.8 kb in size. The relative amount of p53 mRNA in the 16 samples varied over a 27-fold range. No correlation was evident between p53 protein expression (on the basis of the 15-min metabolic labelling assay) and the level of p53 mRNA in AML blasts (Figure 3).

OCI/AML-3 and OCI/AML-4 cells contained similar amounts of p53 protein. However, the RNA blot shown in Figure 2 indicated that the abundance of p53 mRNA was 4-fold higher in OCI/AML-3 than in OCI/AML-4. A 4- to 8-fold difference in p53 RNA was seen in repeated

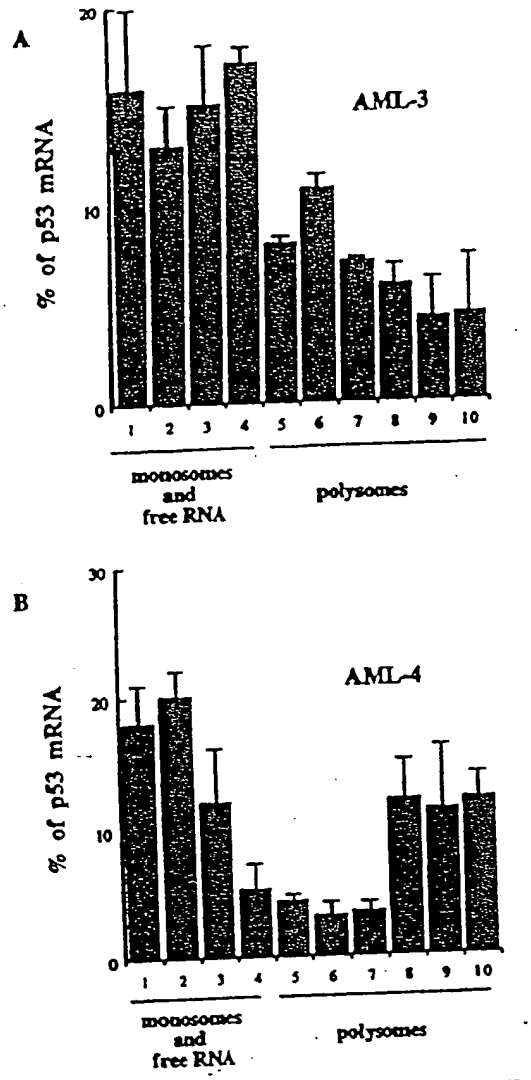


Fig. 4. Association of p53 mRNA with polysomes in OCI/AML-3 (A) and OCI/AML-4 (B) cells. The association of p53 mRNA with polysomes in OCI/AML-3 and OCI/AML-4 cells was compared. Cell extracts containing polysomes were prepared in the presence of cycloheximide and loaded on a 15–50% linear sucrose gradient. Ten fractions were collected and the amount of p53 mRNA in each fraction was determined by dot-blot hybridization analysis with a ^{32}P-labelled human p53 cDNA probe. The size of the polysomes with respect to the gradient was estimated using a polysome preparation from OCI/AML-3 cells. The positions of free ribosomes, monosomes and polysomes are indicated. Error bars represent the standard error of the mean from three separate experiments.

experiments after normalization with probes that detect 18S ribosomal RNA or GAPDH to ensure equivalent loading of RNA samples on the gels. We conclude that p53 RNA levels and p53 protein expression are variable in AML blasts and cell lines, and that the level of p53 protein expression is not related to the amount of p53 mRNA in these cells.

Association of p53 mRNA with polysomes

To test whether p53 gene expression is under translational control *in vivo*, the association of p53 mRNA with polysomes in OCI/AML-3 and OCI/AML-4 cells was analyzed (Figure 4). If p53 mRNA is more translationally active in OCI/AML-4 than in OCI/AML-3 as the above results suggest, then a larger proportion of the p53 mRNA present in OCI/AML-4 should be associated with [illegible] polysomes compared with OCI/AML-3. Cells were collected and lysed in the presence of cycloheximide and $MgCl_2$, which stabilize the association of ribosomes with mRNA. The lysates were sedimented through a linear sucrose gradient and fractions were collected. RNA was extracted from each fraction and analyzed for the presence of p53 mRNA by dot-blot hybridization with a ^{32}P-labelled p53 cDNA probe. The gradients were calibrated with polysomes prepared from lysates by precipitation with 100 mM $MgCl_2$. Polysomes were found at the bottom of the gradient in fractions 5–10, while monosomes were found in fractions 1–4. p53 mRNA from OCI/AML-4 cells was associated with larger polysomes than was p53 mRNA from OCI/AML-3 cells (Figure 4). In OCI/AML-4 cells, 39% of the p53 mRNA was found in fractions 7–10 containing high molecular weight polysomes, while in OCI/AML-3 cells 21% of the p53 mRNA was found in these same fractions. As an internal control, the distribution of ribosomal protein L35 RNA was compared and shown to be identical in OCI/AML-3 and OCI/AML-4 (data not shown).

Analysis of the 5′ end of p53 mRNA

The human p53 gene has been shown to have a cluster of six or seven major transcription initiation sites and several minor sites lying further upstream (Tuck and Crawford, 1989). Transcripts initiating from the minor sites would have a longer 5′UTR with potential to form a stable stem–loop structure close to the 5′ cap. Such structures would not be expected to form in transcripts initiating from the major start sites. 5′-stem–loop structures were described for rodent p53 mRNA (Bienz *et al.*, 1984; Bienz-Tadmor *et al.*, 1985). Recently, mouse p53 protein was shown to bind to the 5′UTR and to inhibit translation of its own mRNA in an *in vitro* assay system (Mosner *et al.*, 1995). Stable stem–loop structures in the 5′UTR regions of a number of mRNA transcripts have been shown to inhibit translation initiation by interfering with the activity of translation initiation factors or by serving as binding sites for regulatory proteins that inhibit translation (Feng and Holland, 1988; Fu *et al.*, 1991; Melefors and Hentze, 1993; Pause *et al.*, 1993).

To determine if the low level of p53 protein expression in leukemic blasts was the result of transcription initiating at the minor start sites, the 5′ ends of p53 mRNA present in different blast samples and cell lines were mapped using an RNase protection assay. A 729 nucleotide antisense RNA probe containing genomic sequences from the p53 promoter region fused with cDNA sequences extending into exon 4 was generated by transcription with SP6 RNA polymerase in the presence of [^{32}P]UTP (Figure 5A). This probe would yield protected p53 fragments of 385 nucleotides corresponding to transcripts originating from the major start site and 449 nucleotides corresponding to transcripts originating from the most 5′ of the minor start sites. Total RNA extracted from OCI/AML-3 and OCI/AML-4 cell lines and from seven AML blast samples was examined. After digestion, the protected fragments were resolved by electrophoresis on a denaturing polyacrylamide gel. As shown in Figure 5B, the predominant protected fragment in all the RNA samples was 385 nucleotides in length indicating a common site for initiation

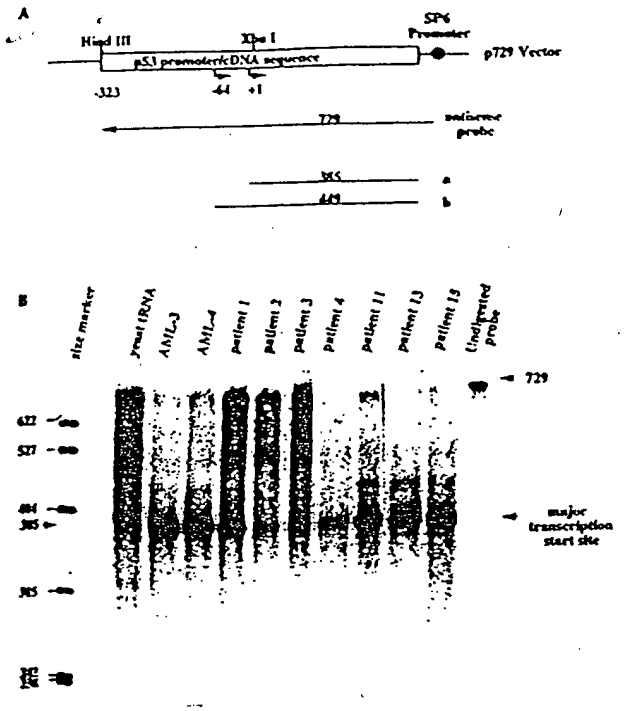


Fig. 5. RNase protection assay. (A) The map of the p729 plasmid. The p729 plasmid was constructed as described under Materials and methods. After linearization with *Hin*dIII, a 729 nucleotide antisense RNA probe was generated by transcription with SP6 RNA polymerase yielding protected p53 fragments of ~385 nucleotides due to p53 transcripts initiating from one of the major start sites (a) and 449 nucleotides due to p53 transcripts initiating from the most 5′ of the minor transcription start sites (b). (B) The 729 nucleotide [^{32}P]UTP-labelled antisense RNA probe was annealed to 30 μg of total RNA extracted from OCI/AML-3 and OCI/AML-4 cell lines and seven AML blast samples before digestion with RNase A and RNase T1. The protected fragments were separated by electrophoresis on a 6% polyacrylamide–8 M urea gel and visualized by autoradiography. The positions and size (nucleotide length) of 5′ end-labelled fragments of *Msp*I-digested pBR322 plasmid DNA are indicated on the left. The bottom arrow indicates the position of the major protected fragment and the top arrow indicates the undigested probe.

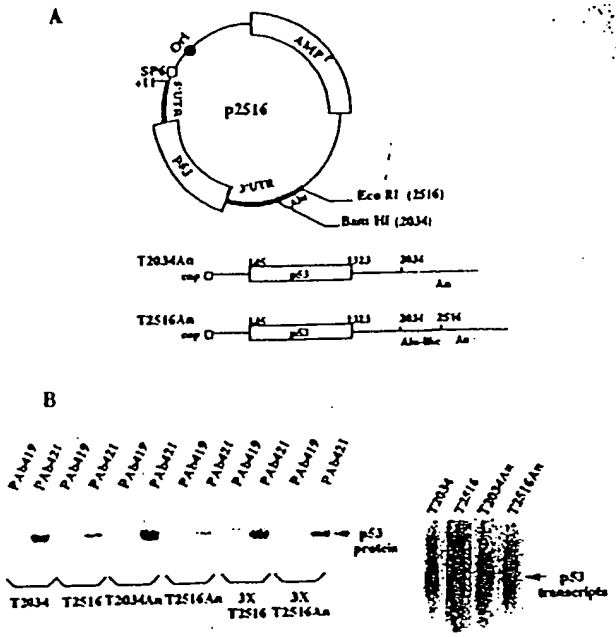


Fig. 6. *In vitro* translation of synthetic p53 RNA containing variable portions of the 3′UTR. (A) Plasmid template used to synthesize p53 RNA *in vitro*. The p2516 plasmid was constructed by inserting the entire 2.5 kb wild-type p53 cDNA sequence downstream of the bacteriophage SP6 promoter in a pSP64-derived plasmid. Transcription from the SP6 promoter present in p2516 leads to the production of transcripts in which the first 10 nucleotides are derived from plasmid sequences while the remaining nucleotides are derived from the p53 gene beginning at the +7 position of native p53 transcripts initiating from one of the major transcription start sites (Tuck and Crawford, 1989). Linearization of p2516 at the *Eco*RI site before *in vitro* transcription generates a full-length, Alu-containing p53 transcript (T2516); linearization at the *Bam*HI site provides a template for the synthesis of a truncated p53 transcript missing a portion of the 3′UTR containing the Alu sequence (T2034). Both transcripts were polyadenylated *in vitro* to generate p2516An and p2034An. The open rectangles shown on the transcripts represent the position of the p53 coding region. (B) 50 ng of the *in vitro*-synthesized T2034, T2516, T2034An and T2516An p53 RNAs were translated in a rabbit reticulocyte lysate at 30°C for 30 min in the presence of [^{35}S]methionine followed by immunoprecipitation, SDS–PAGE and autoradiography. In the 3X T2516 and 3X T2516An lanes, 150 ng of T2516 or T2516An RNA was added to the *in vitro* translation reaction. The right panel presents the results of a Northern blot in which 50 ng of synthetic p53 RNA was applied to an agarose–formaldehyde gel, blotted and hybridized to ^{32}P-labelled human p53 cDNA.

of p53 gene transcription in leukemic blasts at the major start site. These data indicate that, in contrast with murine p53 mRNA, stable secondary structures are unlikely to exist at the 5′ end of human p53 mRNA.

Analysis of the 3′ end of p53 mRNA

Human p53 mRNA contains a long 3′UTR of 1176 nucleotides with an Alu-like repetitive sequence element of ~470 bp located immediately upstream of the poly(A) tail (Matlashewski *et al.*, 1984). The Alu-like sequence is in the reverse transcriptional orientation with respect to the p53 gene. Furthermore, the Alu-like sequence is missing in murine p53 transcripts and it interrupts a region in human p53 mRNA which shows homology to mouse p53 mRNA. When analyzed with the FOLD program of GCG, the Alu-like element in the 3′UTR of human p53 mRNA is predicted to form an independent secondary structure that does not have long-range interactions with other regions of p53 mRNA. In the presence of a poly(A) tail, the secondary structure formed by the Alu-like element is predicted to remain essentially intact except that a 50 nucleotide U-rich sequence at the 5′ boundary of the Alu-like sequence will interact with the poly(A) tail. The extended base pairing between U and A residues will further stabilize the secondary structure formed by the Alu-like element. To determine whether or not the Alu-like repeat present in human p53 mRNA might constitute a negative regulatory element during translation, a series of *in vitro* transcription–translation experiments was performed.

An SP6-derived plasmid containing human wild-type p53 cDNA including the entire 3′UTR was constructed (p2516 in Figure 6A). p2516 was linearized with *Eco*RI or with *Bam*HI and used as a template for *in vitro* transcription. In some reactions, a poly(A) tail of 200–300 adenylic acid residues was added to synthetic p53 RNA using poly(A) polymerase. In this way, four synthetic p53 transcripts were generated: T2516An and T2516 represent full-length, Alu-containing transcripts with or without a poly(A) tail; T2034An and T2034 represent

shorter, Alu-deficient p53 transcripts with or without a poly(A) tail. These transcripts were then used as templates for translation in a rabbit reticulocyte lysate containing [^{35}S]methionine. p53 protein synthesized *in vitro* was immunoprecipitated with PAb421 monoclonal antibody and visualized by autoradiography (Figure 6B). The amount and integrity of the synthetic p53 RNAs added to the *in vitro* translation reactions was monitored by agarose gel electrophoresis and Northern blotting as shown in the right panel of Figure 6B. Densitometric tracing of the data indicated that the Alu-containing, non-polyadenylated transcript T2516 was translated ~3-fold less efficiently than the Alu-deficient, non-polyadenylated transcript T2034. In addition, the polyadenylated, Alu-containing transcript T2516An was translated ~20-fold less efficiently than the polyadenylated, Alu-deficient transcript T2034An. These data indicate that the Alu-like element present in the p53 3′UTR can inhibit p53 mRNA translation *in vitro*, even in the absence of a poly(A) tail. The predicted interaction of the poly(A) tail with the Alu-like element appears to increase further the inhibition of translation.

To test further the inhibitory activity of the p53 3′UTR, we examined the ability of the p53 3′UTR to control the translation of a heterologous RNA. The Alu-containing p53 DNA fragment extending from nucleotides 2034 to 2516 was excised from plasmid p2516 and inserted downstream of a heterologous gene (CAT gene) in an SP6-based plasmid vector to generate the plasmid pCAT-Alu (Figure 7A). *In vitro* transcription and translation revealed that non-polyadenylated CAT-Alu RNA was translated 5-fold less efficiently than non-polyadenylated CAT transcripts lacking the Alu sequence (Figure 7B). When a different region of the p53 3′UTR (nucleotides 1465–2034 in plasmid p2516) with approximately the same length as the Alu-containing fragment was inserted downstream of the CAT gene, no effect on CAT translation was observed (CAT-BS in Figure 7B). The ability of the Alu-containing segment of the p53 3′UTR to act on a heterologous transcript indicates that it likely represses translation independently of upstream sequences.

The inhibitory activity of the Alu-like element on p53 translation was likely the result of its action *in cis* and not simply due to non-specific inhibition of translation, since a 3-fold increase in the amount of Alu-containing transcript added to the reticulocyte lysate resulted in a corresponding increase in the amount of p53 protein synthesized (Figure 6B). Furthermore, when 200 ng of luciferase RNA was added to a reticulocyte lysate together with 200 ng of CAT-Alu or CAT-BS RNA, there was little difference in the amount of luciferase synthesized (Figure 7C). Similarly, when 200 ng of luciferase RNA was added to a reticulocyte lysate, either alone or mixed with 200 ng of T2034 or T2516An RNA, there was little difference in the amount of luciferase synthesized (data not shown).

To confirm that the decrease in p53 protein synthesis from Alu-containing p53 RNAs was due to translational regulation and not due to preferential RNA degradation in the reticulocyte lysate, adenylated T2034 and T2516 synthetic transcripts were added to the rabbit reticulocyte lysate under the same conditions as those used for *in vitro* translation. After incubation for 15 or 60 min, RNA was extracted from the lysate and the amount of synthetic p53 RNA present in the lysate determined by Northern blot analysis. Enhanced degradation of the Alu-containing transcript was not observed (Figure 8). We conclude that a segment of the p53 3′UTR encompassing the Alu-like element is capable of repressing translation *in vitro*.

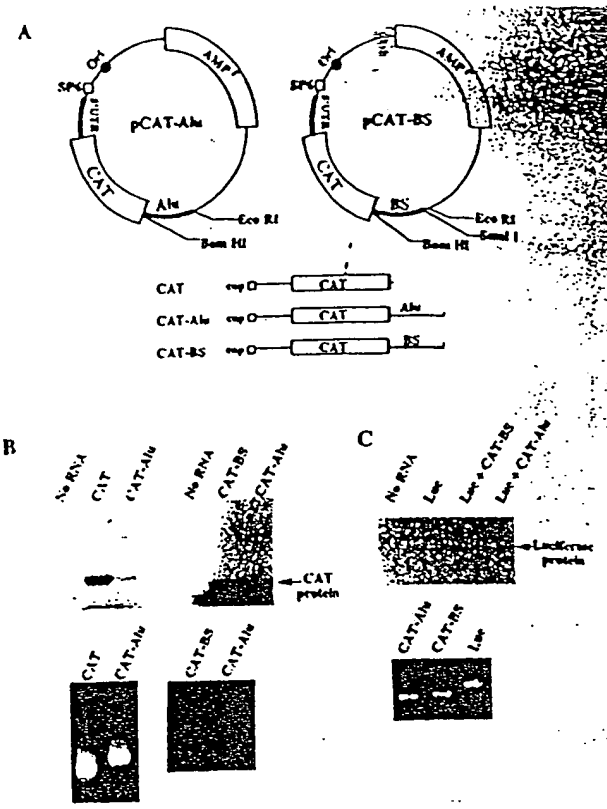


Fig. 7. The p53 Alu-like element can inhibit translation of a heterologous CAT transcript. (A) Plasmids used to generate CAT transcripts *in vitro*. (B) 200 ng of *in vitro*-synthesized CAT, CAT-Alu, and CAT-BS transcripts were translated in a rabbit reticulocyte lysate at 30°C for 30 min in the presence of [^{35}S]methionine. The reactions were stopped by adding an equal volume of the 2X protein sample buffer, heated to 100°C for 5 min and analyzed by SDS–PAGE and autoradiography. An ethidium bromide-stained agarose gel demonstrating the integrity and amount of synthetic transcripts that were added to the *in vitro* translation reaction is shown below. (C) 200 ng of luciferase RNA was translated in a rabbit reticulocyte lysate either alone or in the presence of 200 ng of CAT-BS or 200 ng of CAT-Alu. Reaction mixtures were incubated in the presence of [^{35}S]methionine at 30°C for 30 min and processed as in (B). The *in vitro*-synthesized luciferase protein is shown in the upper panel; the RNA used for *in vitro* translation is shown in the ethidium bromide stained-agarose gel in the bottom panel.

Discussion

The observation that wild-type p53 protein expression in leukemic blast cells does not correlate with the level of p53 mRNA mirrors findings reported previously for blasts and other human cell types (Matlashewski *et al.*, 1986; Kastan *et al.*, 1991a; Slingerland *et al.*, 1991; Sasano *et al.*, 1992; Hsu *et al.*, 1993). The absence of detectable p53 protein in cells expressing abundant levels of wild-type p53 mRNA has usually been attributed to the short half-life of p53 protein in normal cells (Rogel *et al.*, 1985). A similar situation exists in papillomavirus (HPV)-infected cells such as HeLa cells where p53 protein is not detected even though these cells produce p53 mRNA and this RNA is associated with polysomes (Matlashewski

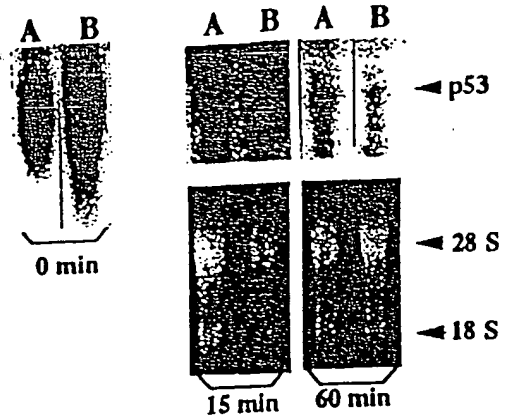


Fig. 8. Stability of synthetic human p53 RNAs in rabbit reticulocyte lysates. 100 ng of adenylated T2034 (A) and T2516 (B) synthetic RNA was added to the rabbit reticulocyte lysate and incubated at 30°C for 15 or 60 min under the same conditions used for *in vitro* translation. RNA present in the lysates was then extracted and loaded on a 1% agarose–formaldehyde gel. The 0 min time point represents 100 ng of synthetic RNA loaded directly on the gel. The amount of p53 RNA in each sample was then determined by Northern blotting using a ^{32}P-labelled human p53 cDNA. The lower panel shows the 28S and 18S ribosomal RNAs recovered from the rabbit reticulocyte lysates detected by ethidium bromide staining of the gel.

et al., 1986). The enhanced degradation of newly synthesized p53 protein in HeLa cells was shown to be promoted by the papillomavirus E6 protein which is expressed constitutively in these cells (Scheffner *et al.*, 1990).

In this report, we present data showing that differences in p53 mRNA abundance exist in AML blasts and that these differences cannot explain the heterogeneity in the level of p53 protein expression in leukemic blast cells. Using a metabolic labelling assay in which blasts from different AML patients were pulse-labelled with [^{35}S]-methionine for 15 min to minimize the contribution of protein half-life on the detection of p53 protein synthesis, we found differences in the level of p53 protein expression in blast samples. These observations raised the possibility that p53 gene expression may be regulated at the translational level in certain human cells. We tested this possibility by analyzing the distribution of p53 mRNA on polysomes *in vivo* and by examining p53 RNA translation *in vitro*.

We have used two AML cell lines, OCI/AML-3 and OCI/AML-4 that contain similar amounts of wild-type p53 protein even though OCI/AML-3 contains 4- to 8-fold more p53 mRNA. Comparison of the polysome profile of these cells indicated that a greater proportion of the p53 mRNA was associated with larger polysomes in OCI/AML-4 than in OCI/AML-3. p53 mRNA in both of these cell lines as well as in blasts from different AML patients is present as a single, full-length species of ~2.8 kb that initiates from a common transcription start site and contains similar sequence and structural elements.

Transcription–translation experiments *in vitro* indicated that the p53 3′UTR contains a negative regulatory domain that is capable of repressing translation *in vitro*. A region of the 3′UTR consisting of ~500 nucleotides and containing an Alu-like element is capable of repressing translation of p53 mRNA and of a heterologous transcript. The p53 3′UTR, when present *in cis*, repressed translation of polyadenylated as well as non-polyadenylated transcripts. Accordingly, we suggest that the Alu-like element, possibly through its secondary structure, is capable of repressing p53 mRNA translation. In addition, interaction of the Alu-like element with the poly(A) tail may repress the latter's function in translation. Experiments are in progress to map precisely this regulatory element in the p53 3′UTR and to determine if the p53 3′UTR plays a similar role in regulating translation *in vivo*.

Our finding that p53 protein expression in AML blasts is controlled, at least in part, through mechanisms acting at the translational level, raises the possibility that translational regulation may provide an epigenetic mechanism to reduce or even eliminate wild-type p53 protein function in leukemic blasts. In preliminary experiments to address this point, we have exposed blast cells that express little or no detectable p53 protein to 6 Gy of ionizing radiation and have observed increased steady-state levels of p53 protein at 1.5 h after irradiation (data not shown). Genotoxic agents have been shown previously to increase the level and/or activity of p53 protein through a post-transcriptional mechanism that is not well understood (Kastan *et al.*, 1991b; Fritsche *et al.*, 1993; Lu and Lane, 1993; Zhan *et al.*, 1993). Hence, blast cells retain the ability to up-regulate p53 expression in response to genotoxic stress. At least under these conditions, p53 function may not be lost. This type of analysis, however, does not address the function of p53 in proliferating cells that have not been exposed to genotoxic stress. In this regard, previous studies from our laboratory demonstrated a highly significant correlation between p53 protein expression in leukemic blast cells and the secondary plating efficiency of these cells (Smith *et al.*, 1986). The latter provides an estimate of the self-renewal capacity of progenitor cells in the blast population. Deregulated p53 expression might, therefore, be expected to affect the self-renewal capacity of blasts in the absence of genotoxic stress.

Accumulating evidence demonstrates the involvement of the 3′UTR in translational control (Jackson, 1993). The demonstration that the 3′UTR of certain transcripts can control mRNA localization and polyadenylation provides a mechanism for translational regulation (Huarte *et al.*, 1992; Gavis and Lehmann, 1994). In addition, specific sequences within 3′UTRs have been shown to repress translation (Goodwin *et al.*, 1993; Evans *et al.*, 1994; Kwon and Hecht, 1993). RNA–protein interactions are likely to be involved in 3′UTR-dependent translational repression. Indeed, a protein that binds specifically to the 3′UTR of protamine 2 mRNA and represses its translation has been identified (Kwon and Hecht, 1993). If the p53 3′UTR can be shown to regulate p53 mRNA translation *in vivo*, it is possible that *trans*-acting factors (missing or inactive in reticulocyte lysates) activate components of the translational machinery to bypass this negative regulatory domain on human p53 mRNA. Such *trans*-acting factors could interact directly with p53 mRNA to enhance its rate of translation. Alternatively, *trans*-acting factors directed to the p53 3′UTR (that are also present in reticulocyte lysates) may act as repressors of translation. Differences in the level of p53 protein synthesis among AML blasts and possibly other human cells could, therefore, be determined by differences in the level or activity of these regulatory molecules.

Materials and methods

Cells

The OCI/AML-3 and OCI/AML-4 cell lines were derived from primary blasts of two AML patients (Wang *et al.*, 1989). The OCI-M2 cell line was derived from the primary blasts of a patient whose erythroleukemia represented the end stage of a previously identified myelodysplastic syndrome (Papayannopoulou *et al.*, 1988). OCI/AML-3 and OCI-M2 cells were grown in alpha-modified minimum essential medium (α-MEM) containing 10% fetal calf serum (FCS) (GIBCO). The OCI/AML-4 cells were grown in α-MEM containing 10% FCS and 10% conditioned medium obtained from the human bladder carcinoma cell line 5637 (5637-CM) (Wang *et al.*, 1989). The AML blast cells were obtained directly from AML patients. The mononuclear cell fraction of peripheral blood was collected after separation through Ficoll–Hypaque (Pharmacia) (1.077 g/ml) and T-lymphocyte depletion (Minden *et al.*, 1979). These cells were stored frozen in liquid nitrogen before use.

Metabolic labelling and immunoprecipitation

The blast cells of AML patients were thawed and incubated for 2 days at 37°C in α-MEM containing 10% FCS and 10% 5637-CM before metabolic labelling. 1×10^7 cells were labelled with 0.2 mCi [^{35}S]-methionine (DuPont NEN Research Products) in 0.5 ml α-MEM lacking methionine and containing 10% dialysed FCS at 37°C for 15 min. Cells were then immediately pelleted, the radioactive medium removed, and the cells lysed on ice in a solution containing 25 mM Tris pH 7.4, 50 mM NaCl, 0.5% sodium deoxycholate, 2% NP40, 0.2% SDS, 0.5 mM phenylmethylsulfonyl fluoride (PMSF), 1 μg/ml leupeptin and 1 μg/ml aprotinin for 20 min. Lysates were cleared by centrifugation, the supernatant was retained and incubated with 5 μg of a non-specific IgG2a mouse monoclonal antibody (Sigma) for 60 min on ice. These were then reacted with 0.5 ml of a 10% suspension of formalin-treated *Staphylococcus aureus* Cowan I cells (Pansorbin, Calbiochem-Behring) for 30 min on ice, followed by centrifugation and retention of the supernatant. Portions of precleared lysates containing equal numbers of trichloroacetic acid-insoluble counts (10^7 c.p.m.) were diluted in NET/GEL buffer (150 mM NaCl, 5 mM EDTA pH 8.0, 50 mM Tris pH 7.4, 0.05% NP40, 0.02% sodium azide, 0.25% gelatin) and immunoprecipitated on ice for 2 h with PAb421 monoclonal antibodies against p53 protein or control PAb419 antibodies (Harlow *et al.*, 1981). The immune complexes were collected on 60 μl prewashed protein A–Sepharose beads (Pharmacia), washed three times with NET/GEL buffer, and eluted into 30 μl protein sample buffer (2% SDS, 10% glycerol, 0.1% bromophenol blue, 25 mM Tris pH 6.8, 0.1 M dithiothreitol) by boiling for 10 min. The Sepharose beads were removed by centrifugation, the samples were loaded on a 10% polyacrylamide gel containing SDS and proteins were resolved by electrophoresis at 45 mA. Gels were fixed in 7.5% acetic acid and 25% methanol for 30 min before drying and exposure to X-ray film (DuPont NEN Research Products).

Western blot analysis

5×10^6 cells were lysed directly in an equal volume of 2X protein sample buffer. The extracts were passed through a 21-gauge needle several times to reduce viscosity and boiled for 10 min before electrophoresis at 45 mA on a 10% polyacrylamide gel containing SDS. Resolved proteins were transferred to a nitrocellulose membrane (Schleicher & Schuell), and the abundance of p53 protein was estimated by immunoblotting with a human p53-specific monoclonal antibody PAb1801 (Banks *et al.*, 1986). Bound antibody was detected using the enhanced chemiluminescence detection system (DuPont NEN Research Products) according to the manufacturer's instructions.

Northern blot analysis

Total cellular RNA was isolated using the guanidinium thiocyanate–cesium chloride method (Chirgwin *et al.*, 1979). 20 μg of total RNA was separated by electrophoresis on a 1% agarose gel containing 6% formaldehyde and transferred to a nitrocellulose membrane (Schleicher & Schuell). The blots were hybridized with cDNA probes labelled with [^{32}P]dCTP in a random priming reaction (Feinberg and Vogelstein, 1983), washed and exposed to X-ray film. The amount of RNA was determined with a Molecular Dynamics PhosphorImager using Multiquant software. The human p53 probe was the *Xba*I–*Eco*RI fragment of p53 cDNA from the pR4-2 plasmid (Harlow *et al.*, 1985); the L35 probe was the *Pst*I–*Bam*HI fragment from the human ribosomal protein L35 cDNA (Herzog *et al.*, 1990); the GAPDH probe was a 1.3 kb *Pst*I fragment of rat GAPDH cDNA (Fort *et al.*, 1985); the 18S ribosomal RNA probe was the *Eco*RI fragment from the human ribosomal RNA gene (Torczynski *et al.*, 1985).

Genomic DNA preparation

Genomic DNA from OCI/AML-3 and OCI/AML-4 cell lines was isolated following a modification of the procedure described by Kupiec *et al.* (1987). 3×10^7 cells were washed with ice-cold PBS buffer, resuspended in 3 ml of lysis buffer (20 mM EDTA pH 8.0, 100 μg/ml proteinase K, 0.5% sarkosyl) and incubated at 50°C for 3 h. DNA was extracted with phenol/chloroform, dialysed against 50 mM Tris–HCl pH 8.0, 10 mM EDTA, 10 mM NaCl at 4°C, and then treated with RNase A (100 μg/ml) at 37°C for 3 h. DNA was again extracted with phenol/chloroform and dialysed against 10 mM Tris pH 7.4, 1 mM EDTA. DNA concentration was determined by measuring the absorbance at 260 nm.

Amplification of p53 sequences from RNA and DNA

20 μg of total RNA was precipitated with ethanol and resuspended in a 30 μl reaction containing 300 ng of oligo(dT) primer (Amersham International), 50 mM Tris–HCl pH 8.3, 77 mM KCl, 3 mM $MgCl_2$, 3 mM dithiothreitol, 3 mM dNTP, 30 units of RNAguard (Pharmacia) and 200 units of Moloney murine leukemia virus reverse transcriptase (GIBCO-BRL) and incubated at 42°C for 60 min. The first strand cDNA was then used as the template for amplification by PCR using *Taq* polymerase (Promega). PCR amplification was performed with 10 μl of each first stand cDNA as the template and 40 cycles of denaturation (94°C, 1 min), annealing (64°C, 30 s), and elongation (72°C, 1 min). The following p53-specific primers were used for amplifying the complete coding region and the 3'UTR: 5'SX1 (sense, exon 1, GACACTTT-GCGTTCGGGCTGGGAG), 5'SX5A (sense, exon 5, GAGCGCTGCT-CAGATAGCGATG), 3'SX11 (sense, exon 11, GAAGGGCCTGACT-CAGACTGAC), 3'AX-6 (antisense, exon 6, AGATGCTGAGGAGGG-GCCAGAC), JS-3 (antisense, exon 11, GAGGGAGAGATGGGGGT-GGGAGGCTGTC) and AS-4 (antisense, exon 11, GGCAGCAAAGT-TTTATTGTAAAATAAG). The 5'UTR and sequences further upstream were amplified from 1 μg genomic DNA using the following pair of p53-specific primers: 5'UTR-1 (sense, promoter region, ACCTAA-GCTTGTCATGGCGACTGTCCAGCTTTG) and p-EX (antisense, exon 1, CCAATCCAGGGAAGCGTGTCACCG).

Direct sequencing of double-stranded PCR products

Double-stranded DNA fragments produced by PCR amplification were eluted from agarose gels and purified by extraction with phenol/chloroform. 200 ng of purified PCR product were mixed with human p53-specific oligonucleotides as sequencing primers, frozen in dry ice, dried in a centrifugal evaporator (Savant SpeedVac), redissolved in sequencing buffer (40 mM Tris–HCl pH 7.5, 25 mM $MgCl_2$, 50 mM NaCl, 10% DMSO) and subjected to the sequencing reaction as described by Winship (1989).

RNase protection assay

Plasmid p729 was constructed from three DNA fragments in two stages. A 330 bp DNA fragment derived from the human p53 gene promoter was excised from the p2E-H2BX plasmid (Lamb and Crawford, 1986) with *Hin*dIII and *Xba*I and inserted into the pGEM-4 plasmid (Promega) between the *Hin*dIII and *Xba*I sites. In the second stage, a fragment corresponding to the 5' end of p53 mRNA was obtained by RT–PCR using p53 mRNA prepared from OCI/AML-3 cells and the p53-specific primers 5'UTR-3 (sense, exon 1, CCGGAAGCTTCAAAAGTCTA-GAGCCACCGTCCAG) and 5'AX4 (antisense, exon 4, GGTGTAGG-AGCTGCTGCTGGTGC). The resulting fragment was end-filled with the Klenow fragment of DNA polymerase I, digested with *Xba*I at the site present in the 5'UTR-3 primer shown underlined and inserted between the *Xba*I and *Sma*I sites present in the plasmid generated in the first stage.

p729 was linearized with *Hin*dIII and a 729 nucleotide antisense probe was prepared by transcription with SP6 RNA polymerase. The *in vitro* transcription reaction mixture contained 50 mM Tris–HCl pH 8.0, 10 mM $MgCl_2$, 4 mM spermidine, 10 mM NaCl, 0.5 mM each of ATP, GTP, CTP, 12 μM UTP, 5 μCi [^{32}P]UTP, 10 mM dithiothreitol, 20 units of RNAguard, 0.5 μg of linearized template and 10 units of SP6 RNA polymerase in a final volume of 20 μl. After incubation at 37°C for 60 min, the DNA template was digested with DNase I and the RNA probe was extracted with phenol/chloroform, precipitated with ethanol and resuspended in water. This RNA probe covered the entire p53 gene promoter region and included the first three exons and a part of the fourth exon. p53 transcripts initiating from one of the major start sites

should yield protected fragments of ~385 nucleotides. p53 transcripts originating from the most 5' of the minor start sites should yield protected fragments of 449 nucleotides (Tuck and Crawford, 1989).

In the RNase protection assay, 30 µg of total RNA was mixed with 1×10^5 c.p.m. of the labelled probe and precipitated with ethanol. The RNA/probe mixtures were then washed, dried and resuspended in 10 µl of hybridization solution (Winter *et al.*, 1985), heated to 80°C for 10 min, and hybridized at 46°C overnight. After hybridization, the samples were mixed with 0.18 ml of RNase digestion mix containing 60 µg/ml of RNase A (type III, Sigma), 1100 U/ml of RNase T1 (Boehringer Mannheim) in 300 mM NaCl, 5 mM EDTA, 10 mM Tris–HCl pH 7.5. After incubation at 37°C for 60 min, the digestion was terminated by addition of 10 µl of 20% SDS and 5 µl of proteinase K (10 mg/ml) (Boehringer Mannheim) and incubation at 37°C for 15 min. Protected fragments were extracted with phenol/chloroform, precipitated with ethanol, resolved by denaturing gel electrophoresis and visualized by autoradiography.

Polysome analysis

5×10^7 cells were washed once in ice-cold Tris–saline solution (25 mM Tris–HCl pH 7.5, 25 mM NaCl) containing 10 mM $MgCl_2$ and 10 µg/ml cycloheximide. The cells were then immediately lysed on ice with the use of a Dounce homogenizer in 2 ml homogenization buffer containing 25 mM Tris–HCl pH 7.5, 25 mM NaCl, 10 mM $MgCl_2$, 2% Triton X-100, 340 U/ml heparin (LEO Laboratories Canada Ltd), 2 mM vanadyl ribonucleoside complex (Sigma), 2.5 mM PMSF, 10 µg/ml cycloheximide, 1 mM dithiothreitol and 1 mM EGTA. The extract was centrifuged at 14 000 r.p.m. for 6 min at 4°C to remove cell debris, the supernatant was collected and layered over a 15–50% linear sucrose gradient (11 ml) prepared in homogenization buffer. The gradients were centrifuged in an SW41 Beckman rotor at 175 000 *g* for 110 min at 4°C. Ten fractions of equal volume were collected from the bottom of the tubes. RNA was prepared from each of the fractions by phenol/chloroform extraction and ethanol precipitation and resuspended in 200 µl DEPC-treated water. The amount of p53 mRNA in each fraction (100 µl of the RNA sample) was determined by dot-blot hybridization analysis using a ^{32}P-labelled human p53 cDNA probe. Polysomes used to calibrate the gradients were prepared in exactly the same way except for an additional purification step involving precipitation of the polysomes present in the homogenate with 100 mM $MgCl_2$ for 1 h on ice before sucrose gradient sedimentation. For calibration, 0.3-ml fractions were collected from the bottom of the gradient and A_{254} of each fraction was determined.

Templates for in vitro transcription and translation

Plasmid p2516 contains nearly full-length human wild-type p53 cDNA and was constructed by the correct ligation of three cDNA fragments. One fragment corresponding to the 5' end of the p53 transcript was obtained from pR4-2 (Harlow *et al.*, 1985) after digestion with *Xba*I and *Pvu*II which cut in exons 1 and 5, respectively. The middle fragment was obtained from pProSp53 (Matlashewski *et al.*, 1987) after digestion with *Pvu*II and *Bam*HI which cut in exons 5 and 11, respectively. The third fragment corresponding to the 3' end of the p53 transcript was obtained by RT–PCR amplification of the 3'UTR of p53 mRNA using p53-specific oligonucleotides as primers, 3'SX13 (sense, exon 11, GTCACCCCATCCCACACCCTGG) and AS-4. The PCR-amplified fragment was end-filled with the Klenow fragment of DNA polymerase I and digested at an internal *Bam*HI site. These three fragments which represent contiguous sequences of the native p53 transcript were inserted between the *Xba*I and *Sma*I sites of a modified form of the pSP64 vector (Promega) in which polylinker sequences between the *Hin*dIII site and the *Xba*I site were deleted. The resulting plasmid is referred to as p2516 and yields a p53 transcript *in vitro* starting with the sequence 5'GAATACAAGCTCTAGA.....3'. The *in vitro* transcript is nearly identical to p53 transcripts originating from the most 3' of the major transcription initiation sites *in vivo* which start with 5'CAAAAGTCTAGA.....3' (Tuck and Crawford, 1989). The beginning of identity corresponding to an *Xba*I site in the cDNA is underlined. Digestion of p2516 with *Eco*RI provides a template that can produce a synthetic full-length p53 transcript of 2516 nucleotides. Digestion with *Bam*HI provides a template for a truncated p53 transcript of 2034 nucleotides that is missing sequences from the 3'UTR containing the Alu-like element.

The plasmid pCAT-Alu was constructed in two steps. First, the chloramphenicol acetyltransferase gene was excised from the CAT plasmid (Fu *et al.*, 1991) with *Hin*dIII and *Bam*HI, and inserted into pSP64 to generate pSP6CAT. Second, the *Bam*HI–*Eco*RI fragment from p2516 that contains the Alu-like element present in the p53 3'UTR was inserted immediately downstream of the CAT gene. The plasmid pCAT-BS was constructed by removing the *Sma*I–*Bam*HI fragment of the p53 3'UTR present in p2516 and inserting this fragment in reverse orientation into pSP6CAT immediately downstream of CAT. This *Sma*I–*Bam*HI fragment is missing the Alu-like element present at the distal end of the p53 3'UTR.

In vitro transcription and in vitro polyadenylation

Plasmid DNAs containing templates for *in vitro* transcription were linearized at selected restriction endonuclease sites. Standard transcription assays (Melton *et al.*, 1984) were performed as described above for the preparation of antisense RNA probes with the omission of [^{32}P]UTP. 0.5 mM ^{7m}G(5')ppp(5')G and 0.05 mM GTP were included in the reactions to provide efficient capping at the 5' end of synthetic transcripts. Polyadenylation reactions contained synthetic RNA, 0.2 mM ATP, 50 mM Tris–HCl pH 8.0, 10 mM $MgCl_2$, 250 mM NaCl, 2 mM $MnCl_2$, 2 mM dithiothreitol, 1 unit/µl RNAguard (Pharmacia), 500 µg/ml of BSA (Pharmacia) and 5 units of poly(A) polymerase (Pharmacia) in a 50 µl final volume (McGrew *et al.*, 1989). After 30 min at 37°C, polyadenylated RNAs were purified by phenol/chloroform extraction and ethanol precipitation.

In vitro translation and immunoprecipitation

Synthetic transcripts were translated in micrococcal-nuclease-treated rabbit reticulocyte lysates (Promega) under the conditions recommended by the supplier. Reactions containing p53 transcripts were incubated for 30 min at 30°C in the presence of [^{35}S]methionine and stopped by addition of dithiothreitol to a final concentration of 1 mM and EDTA pH 8.0 to a final concentration of 10 mM. Each reaction was then divided into two aliquots, one for immunoprecipitation with the p53-specific monoclonal antibody PAb421 and the other for immunoprecipitation with a control antibody PAb419. Reactions containing CAT or luciferase transcripts were incubated for 30 min at 30°C in the presence of [^{35}S]methionine and were stopped by addition of protein sample buffer, boiled for 5 min and resolved by polyacrylamide gel electrophoresis.

Acknowledgements

This work was supported by grants from the Medical Research Council of Canada and from the National Cancer Institute of Canada.

References

Banks,L., Matlashewski,G. and Crawford,L. (1986) Isolation of human p53-specific monoclonal antibodies and their use in the studies of human p53 expression. *Eur. J. Biochem.*, 159, 529–534.

Benchimol,S., Munroe,D.G., Peacock,J., Gray,D. and Smith,L.J. (1989) Abnormalities in structure and expression of the p53 gene in leukemia. *Cancer Cells*, 7, 121–125.

Bienz,B., Zakut-Houri,R., Givol,D. and Oren,M. (1984) Analysis of the gene coding for the murine cellular tumour antigen p53. *EMBO J.*, 3, 2179–2183.

Bienz-Tadmor,B., Zakut-Houri,R., Libresco,S., Givol,D. and Oren,M. (1985) The 5' region of the p53 gene: evolutionary conservation and evidence for a negative regulatory element. *EMBO J.*, 4, 3209–3213.

Chirgwin,J.M., Przybyla,A.E., McDonald,R.J. and Rutter,W.J. (1979) Isolation of biochemically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry*, 18, 5294–5299.

Evans,T.C., Crittenden,S.L., Kodoyianni,V. and Kimble,J. (1994) Translational control of maternal *glp-1* mRNA establishes an asymmetry in the *C. elegans* embryos. *Cell*, 77, 183–194.

Feinberg,A.P. and Vogelstein,B. (1983) A technique for radio-labeling DNA restriction endonuclease fragments to high specific activities. *Anal. Biochem.*, 132, 6–13.

Fenaux,P., Jonveaux,P., Quiquandon,I., Laï,J.L., Pignon,J.M., Loucheux-Lefebvre,M.H., Bauters,F., Berger,R. and Kerckaert,J.P. (1991) p53 gene mutations in acute myeloid leukemia with 17p monosomy. *Blood*, 78, 1652–1657.

Fenaux,P., Preudhomme,C., Quiquandon,I., Jonveaux,P., Laï,J.L., Vanrumbeke,M., Loucheux-Lefebvre,M.H., Bauters,F., Berger,R. and Kerckaert,J.P. (1992) Mutations of the p53 gene in acute myeloid leukaemia. *J. Haematol.*, 80, 178–183.

Feng,S. and Holland,F.C. (1988) HIV-1 *tat* trans-activation requires the loop sequence within *tat*. *Nature*, 334, 165–167.

Fialkow,P.J., Singer,J.W., Raskind,W.H., Adamson,J.W., Jacobson,R.J., Bernstein,I.D., Dow,L.W., Najfeld,V. and Veith,R. (1987) Clonal

development, stem-cell differentiation, and clinical remission in acute nonlymphocytic leukemia. *N. Engl. J. Med.*, 317, 468–473.

Fritsche,M., Haessler,C. and Brandner,G. (1993) Induction of nuclear accumulation of the tumor-suppressing protein p53 by DNA-damaging agents. *Oncogene*, 8, 307–318.

Fort,P., Marty,L., Piechaczyk,M., Sabrouty,S.E., Dani,C., Jeanteur,P. and Blanchard,J.M. (1985) Various rat adult tissues express only one major mRNA species from the glyceraldehyde-3-phosphate-dehydrogenase multigenic family. *Nucleic Acids Res.*, 13, 1431–1442.

Fu,L., Ye,R., Browder,L.W. and Johnston,R.N. (1991) Translational potentiation of messenger RNA with secondary structure in *Xenopus*. *Science*, 251, 807–810.

Gavis,E.R. and Lehmann,R. (1994) Translational regulation of *nanos* by RNA localization. *Nature*, 369, 315–318.

Goodwin,E.B., Okkema,P.G., Evans,T.C. and Kimble,J. (1993) Translational regulation of *tra-2* by its 3' UTR controls sexual identity in *C. elegans*. *Cell*, 75, 329–339.

Harlow,E., Crawford,L.V., Pim,D.C. and Williamson,N.M. (1981) Monoclonal antibodies specific for Simian virus 40 tumor antigen. *J. Virol.*, 39, 861–869.

Harlow,E., Williamson,N.M., Ralston,R., Helfman,D.M. and Adams,T.E. (1985) Molecular cloning and *in vitro* expression of a DNA clone for human cellular tumor antigen p53. *Mol. Cell. Biol.*, 5, 1601–1610.

Herzog,H., Hofferer,L., Schneider,R. and Schweiger,M. (1990) cDNA encoding the human homologue of rat ribosomal protein L35a. *Nucleic Acids Res.*, 18, 4600.

Hsu,H.C., Tseng,H.J., Lai,P.L., Lee,P.H. and Peng,S.Y. (1993) Expression of p53 gene in 184 unifocal hepatocellular carcinomas: association with tumor growth and invasiveness. *Cancer Res.*, 53, 4691–4694.

Huarte,J., Stutz,A., O'Connell,M.L., Guber,P., Belin,D., Darrow,A.L., Strickland,S. and Vassalli,J.D. (1992) Transient translational silencing by reversible mRNA deadenylation. *Cell*, 69, 1021–1030.

Jackson,R.J. (1993) Cytoplasmic regulation of mRNA function: the importance of the 3' UTR. *Cell*, 74, 9–14.

Kastan,M.B. *et al.* (1991a) Levels of p53 protein increase with maturation in human hematopoietic cells. *Cancer Res.*, 51, 4279–4286.

Kastan,M.B., Onyekwere,O., Sidransky,D., Vogelstein,B. and Craig,R.W. (1991b) Participation of p53 protein in the cellular response to DNA damage. *Cancer Res.*, 51, 6304–6311.

Kupiec,J.J., Giron,M.L., Viletter,D., Jeltsch,J.M. and Emanoil-Ravier,R. (1987) Isolation of high-molecular-weight DNA from eukaryotic cells by formamide treatment and dialysis. *Anal. Biochem.*, 164, 53–59.

Kwon,Y.K. and Hecht,N.B. (1993) Binding of a phosphoprotein to the 3' untranslated region of the mouse protamine 2 mRNA temporally represses its translation. *Mol. Cell. Biol.*, 13, 6547–6557.

Lai,J.L., Preudhomme,C., Zandecki,M., Flactif,M., Vanrumbeke,M., Lepelley,P., Wattel,E. and Fenaux,P. (1995) Myelodysplastic syndromes and acute myeloid leukemia with 17p deletion. An entity characterized by specific dysgranulopoiesis and a high incidence of P53 mutations. *Leukemia*, 9, 370–381.

Lamb,P. and Crawford,L. (1986) Characterization of the human p53 gene. *Mol. Cell. Biol.*, 6, 1379–1385.

Lu,X. and Lane,D.P. (1993) Differential induction of transcriptionally active p53 following UV or ionizing radiation: Defects in chromosome instability syndromes? *Cell*, 75, 765–778.

Matlashewski,G., Lamb,P., Pim,D., Peacock,J., Crawford,L. and Benchimol,S. (1984) Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene. *EMBO J.*, 13, 3257–3262.

Matlashewski,G., Banks,L., Pim,D. and Crawford,L.V. (1986) Analysis of human p53 proteins and mRNA levels in normal and transformed cells. *Eur. J. Biochem.*, 154, 665–672.

Matlashewski,G.J., Tuck,S., Pim,D., Lamb,P., Schneider,J. and Crawford, L.V. (1987) Primary structure polymorphism at amino residue 72 of human p53. *Mol. Cell. Biol.*, 7, 961–963.

Melefors,O. and Hentze,M.W. (1993) Translational regulation by mRNA–protein interactions in eukaryotic cells: ferritin and beyond. *BioEssays*, 15, 85–90.

Melton,D.A., Krieg,P., Rebagliati,M., Maniatis,T., Zinn,K. and Green,M. (1984) Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Res.*, 12, 7035–7056.

Milner,J. (1991) A conformation hypothesis for the suppressor and promoter functions of p53 in cell growth control and in cancer. *Proc. R. Soc. Lond. B.*, 245, 139–145.

Minden,M.D., Buick,R.N. and McCulloch,E.A. (1979) Separation of blast cell and T-lymphocyte progenitors in the blood of patients with acute myeloblastic leukemia. *Blood*, 54, 186–195.

Moll,U.M., Riou,G. and Levine,A.J. (1992) Two distinct mechanisms alter p53 in breast cancer: mutation and nuclear exclusion. *Proc. Natl Acad. Sci. USA*, 89, 7262–7266.

Moll,U.M., LaQuaglia,M., Benard,J. and Riou,G. (1995) Wild-type p53 protein undergoes cytoplasmic sequestration in undifferentiated neuroblastomas but not in differentiated tumors. *Proc. Natl Acad. Sci. USA*, 92, 4407–4411.

Momand,J., Zambetti,G.P., Olson,D.C., George,D. and Levine,A.J. (1992) The *mdm-2* oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell*, 69, 1237–1245.

Mosner,J., Mummenbrauer,T., Bauer,C., Sczakiel,G., Grosse,F. and Deppert,W. (1995) Negative feedback regulation of wild-type p53 biosynthesis. *EMBO J.*, 12, 4739–4746.

Oliner,J.D., Kinzler,K.W., Meltzer,P.S., George,D and Vogelstein,B. (1992) Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature*, 358, 80–83.

Papayannopoulou,T., Nakamoto,B., Kurachi,S., Tweeddale,M. and Messner,H. (1988) Surface antigenic profile and globin phenotype of two new human erythroleukemia lines: characterization and interpretations. *Blood*, 72, 1029–1038.

Pause,A., Methot,N. and Sonenberg,N. (1993) The HRIGRXXR region of the DEAD box RNA helicase eukaryotic translation initiation factor 4A is required for RNA binding and ATP hydrolysis. *Mol. Cell. Biol.*, 13, 6789–6798.

Rogel,A., Popliker,M., Webb,C.G. and Oren,M. (1985) p53 cellular tumor antigen: analysis of mRNA levels in normal adult tissues, embryos, and tumors. *Mol. Cell. Biol.*, 5, 2851–2855.

Sasano,H., Goukon,Y., Nishihira,T. and Nagura,H. (1992) *In situ* hybridization and immunohistochemistry of p53 tumor suppressor gene in human esophageal carcinoma. *Am. J. Pathol.*, 141, 545–550.

Scheffner,M., Werness,B.A., Huibregtse,J.M., Levine,A.J. and Howley, P.M. (1990) The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell*, 63, 1129–1136.

Slingerland,J.M., Minden,M.D. and Benchimol,S. (1991) Mutations of the p53 gene in human acute myelogenous leukemia. *Blood*, 7, 1500–1507.

Smith,L.J., McCulloch,E.A. and Benchimol,S. (1986) Expression of the p53 oncogene in acute myeloblastic leukemia. *J. Exp. Med.*, 164, 751–761.

Sugimoto,K., Toyoshima,H., Sakai,R., Miyagawa,K., Hagiwara,K., Hirai,H., Ishikana,F. and Takaku,F. (1991) Mutations of the p53 gene in lymphoid leukemia. *Blood*, 77, 1153–1156.

Sugimoto,K., Hirano,N., Toyoshima,H., Chiba,S., Mano,H., Takaku,F., Yazaki,Y. and Hirai,H. (1993) Mutations of the p53 gene in myelodysplastic syndromes and MDS-derived leukemia. *Blood*, 81, 3022–3026.

Torczynski,R.M., Fuke,M. and Bollon,A.P. (1985) Cloning and sequencing of a human 18S ribosomal RNA gene. *DNA*, 4, 283–291.

Trecca,D., Longo,L., Biondi,A., Cro,L., Calori,R., Grignani,F., Maiolo,A.T., Pelicci,P.G. and Neri,A. (1994) Analysis of p53 gene mutations in acute myeloid leukemia. *Am. J. Hematol.*, 46, 304–309.

Tuck,S.P. and Crawford,L. (1989) Characterization of the human p53 gene promoter. *Mol. Cell. Biol.*, 9, 2163–2172.

Ueda,H., Ullrich,S.J., Gangemi,J.D., Kappel,C.A., Ngo,L., Feitelson, M.A. and Jay,G. (1995) Functional inactivation but not structural mutation of p53 causes liver cancer. *Nature Genet.*, 9, 41–47.

Ullrich,S.J., Mercer,W.E. and Appella,E. (1992) Human wild-type p53 adopts a unique conformational and phosphorylation state *in vivo* during growth arrest of glioblastoma cells. *Oncogene*, 7, 1635–1643.

Wang,C., Curtis,J.E., Minden,M.D. and McCulloch,E.A. (1989) Expression of a retinoic acid receptor gene in myeloid leukemia cells. *Leukemia*, 3, 264–269.

Wattel,E., Preudhomme,C., Hecquet,B., Vanrumbeke,M., Quesnel,B., Dervite,I., Morel,P. and Fenaux,P. (1994) p53 mutations are associated with resistance to chemotherapy and short survival in hematologic malignancies. *Blood*, 84, 3148–3157.

Wiggans,R.G., Jacobson,R.J., Fialkow,P.J., Woolley,P.V., Macdonald,S.J. and Schein,P.S. (1978) Probable clonal origin of acute myeloblastic leukemia following radiation and chemotherapy of colon cancer. *Blood*, 52, 650–663.

Winship,P.R. (1989) An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. *Nucleic Acids Res.*, 17, 1266.

Winter.E., Yamamoto,F., Almoguera.C. and Perucho.M. (1985) A method to detect and characterize point mutations in transcribed genes: amplification and overexpression of the mutant C-Ki-ras allele in human tumor cells. *Proc. Natl Acad. Sci. USA*, 82, 7575–7579.

Zhan.Q., Carrier.F. and Fornace.A.J.Jr (1993) Induction of cellular p53 activity by DNA-damaging agents and growth arrest. *Mol. Cell. Biol.*, 13, 4242–4250.

Zhang.W., Hu.G., Estey.E., HesterJ. and Deisseroth.A. (1992) Altered conformation of the p53 protein in myeloid leukemia cells and mitogen-stimulated normal blood cells. *Oncogene*, 7, 1645–1647.

Zhu.Y.M., Bradbury.D. and Russell.N. (1993) Expression of different conformations of p53 in the blast cells of acute myeloblastic leukaemia is related to *in vitro* growth characteristics. *Br. J. Cancer*, 68, 851–855.

Received on November 30, 1995; revised on April 18, 1996

Degradation of the E7 human papillomavirus oncoprotein by the ubiquitin-proteasome system: targeting via ubiquitination of the N-terminal residue

Eyal Reinstein[1], Martin Scheffner[2], Moshe Oren[3], Aaron Ciechanover*[,1] and Alan Schwartz[4]

[1]*Department of Biochemistry and the Rappaport Family Institute for Research in the Medical Sciences, The Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 31096, Israel;* [2]*Institut für Biochemie, Medizinische Fakultaet Universität zu Koeln, 50931 Koeln, Germany;* [3]*Department of Molecular Cell Biology, The Weizmann Institute of Science, Rehovot 76100, Israel;* [4]*Departments of Pediatrics and of Molecular Biology and Pharmacology, Washington University School of Medicine and St. Louis Children's Hospital, St. Louis, Missouri, MO 63110-1093, USA*

The E7 oncoprotein of the high risk human papillomavirus type 16 (HPV-16), which is etiologically associated with uterine cervical cancer, is a potent immortalizing and transforming agent. It probably exerts its oncogenic functions by interacting and altering the normal activity of cell cycle control proteins such as $p21^{WAF1}$, $p27^{KIP1}$ and pRb, transcriptional activators such as TBP and AP-1, and metabolic regulators such as M2-pyruvate kinase (M2-PK). Here we show that E7 is a short-lived protein and its degradation both *in vitro* and *in vivo* is mediated by the ubiquitin-proteasome pathway. Interestingly, ubiquitin does not attach to any of the two internal Lysine residues of E7. Substitution of these residues with Arg does not affect the ability of the protein to be conjugated and degraded; in contrast, addition of a Myc tag to the N-terminal but not to the C-terminal residue, stabilizes the protein. Also, deletion of the first 11 amino acid residues stabilizes the protein in cells. Taken together, these findings strongly suggest that, like MyoD and the Epstein Barr Virus (EBV) transforming Latent Membrane Protein 1 (LMP1), the first ubiquitin moiety is attached linearly to the free N-terminal residue of E7. Additional ubiquitin moieties are then attached to an internal Lys residue of the previously conjugated molecule. The involvement of E7 in many diverse and apparently unrelated processes requires tight regulation of its function and cellular level, which is controlled in this case by ubiquitin-mediated proteolysis. *Oncogene* (2000) **19**, 5944–5950.

Keywords: human papilloma-virus (HPV); E7; ubiquitin; proteolysis; N-terminus

Introduction

The E7 oncoprotein of the high risk human papillomavirus type 16, which is etiologically associated with pathogenesis of human uterine cervical cancer, is a potent immortalizing and transforming protein. Expression of E7 can transform rodent fibroblasts (Kanda *et al.*, 1988), and in conjunction with an activated Ras oncogene, primary rodent cells (Phelps *et al.*, 1988). Continued expression of the E7 gene is required for the maintenance of the transformed phenotype (Crook *et al.*, 1989), and expression of the protein in non-metastatic mouse cell lines renders the cells metastatic in nude mice (Chen *et al.*, 1993). In transgenic mice, co-expression of E7 along with E6, another high risk HPV oncoprotein, elicits epidermal hyperplasia (Auewarakul *et al.*, 1994), verrucose lesions and papillomas (Greenhalgh *et al.*, 1994). Furthermore, E6 and E7 can cooperate to induce various tumors when expressed ectopically in transgenic mice (Arbeit *et al.*, 1993; Pan and Griep, 1994). Finally, both E7 and E6 are necessary and sufficient to immortalize their primary host cells, human squamous epithelial cells (Hawley-Nelson *et al.*, 1989; Munger *et al.*, 1989).

While the molecular mechanisms that underlie the transforming and immortalizing activity of E7 are still obscure, the protein appears to exert most of its oncogenic functions by interacting physically with key cellular regulatory proteins which leads to modulation of their normal activity. One main function of E7 is its ability to deregulate control of cell cycle progression, allowing cells to exit G0 and enter S phase. It has been shown that via its cd2 domain, E7 binds to the cell cycle regulators p107, p130 and pRb (Arroyo *et al.*, 1993; Davies *et al.*, 1993; Hu *et al.*, 1995). Normally, these proteins function as transcriptional repressors that lead to G1 arrest. It was suggested that the binding of E7 to these proteins leads to their dissociation from their complex with E2F which correlates with stimulation of E2F-dependent transcription. It has also been shown that E7 interacts with both $p27^{KIP1}$ (Zerfass-Thome *et al.*, 1996) and $p21^{WAF1}$ (Funk *et al.*, 1997). Consequently, both proteins fail to block the activity of Cyclin E/Cdk2 complexes which allow transition of the cell across the G1/S border. Binding of E7 to the Jun component of AP-1 can lead to activation of AP-1 driven genes (Antinore *et al.*, 1996). It has been also shown that E7 binds to M2-pyruvate kinase (M2-PK), lowers its affinity to phosphoenol-pyruvate, and thus slows influx of substrates into the tricarboxylic, citric acid cycle (Zweschke *et al.*, 1999). This leads to accumulation of upstream phosphometabolites which serve as precursors to amino acids and nucleotides. The pool of these precursors is low in resting cells, but its expansion is necessary during rapid cell division. E7 can act however in a different mechanism; similar to targeting of p53 for degradation by HPV E6 (Scheffner *et al.*, 1990), it has been shown that association of E7 with pRb also targets the repressor for ubiquitin-mediated degradation (Boyer *et al.*, 1996). Targeting of pRb, and potentially of other regulatory proteins, for degradation, may serve as a second mechanism, besides

*Correspondence: A Ciechanover, Department of Biochemistry, Faculty of Medicine, Technion-Israel Institute of Technology, PO Box 9649, Efron Street, Bat Galim, Haifa 30196, Israel

Received 17 July 2000; revised 21 August 2000; accepted 3 October 2000

physical interaction, by which E7 exerts its deregulatory effects. In the case of pRb, removal of the protein induces the activity of the E2F family of cellular transcription factors which are known to control the expression of the major cell cycle regulatory genes at the G1/S transition.

It has been reported that E7 is short-lived (Selvey *et al*., 1994), however the system involved in its degradation and the mechanism(s) that underlie the process have remained obscure. Many studies have implicated the ubiquitin pathway in the degradation of various short-lived key regulatory proteins. It is involved in proteolysis and processing of many cellular proteins, including cell cycle regulators, oncoproteins and tumor suppressors, transcriptional activators, ER membrane proteins and cell surface receptors. In most cases, ubiquitination of the target protein signals its degradation by the 26S proteasome. Degradation of a protein via the ubiquitin-proteasome pathway involves two discrete and successive steps: (i) covalent attachment of multiple ubiquitin molecules to the protein substrate; and (ii) degradation of the tagged protein by the 26S proteasome. Conjugation of ubiquitin involves activation and transfer of ubiquitin from the ubiquitin-activating enzyme, E1, to one of several ubiquitin-carrier proteins, E2s (known also as ubiquitin-conjugating enzymes, UBCs). E2 transfers the activated ubiquitin moiety to the target substrate that is specifically bound to a member of the ubiquitin-protein ligase family, E3. Subsequent processive transfer of additional activated ubiquitin molecules and their conjugation to previously attached moieties, generates a polyubiquitin chain that serves as a degradation signal for the proteasome. Binding of the substrate to E3 plays an essential role in specific substrate recognition (for recent reviews on the ubiquitin-proteasome pathway, see, for example, Kornitzer and Ciechanover, 2000; Voges *et al*., 1999). In most cases, the first ubiquitin molecule is transferred to an ε-NH_2 group of an internal Lysine residue of the target protein. However, targeting of the myogenic transcription factor MyoD (Breitschopf *et al*., 1998) and of the Epstein Barr Virus (EBV) Latent Membrane Protein-1 (LMP-1; Aviel *et al*., 2000) involves initial ubiquitination of the N-terminal residue followed by synthesis of a poly-ubiquitin chain attached to an internal Lys residue of this N-terminally attached ubiquitin moiety. Thus, unlike many known substrates of the ubiquitin system, degradation of these proteins does not require any internal Lys residue. It has also been reported that a mutant lysine-less α chain of the T cell receptor (TCR) is also degraded by the proteasome, in a process that depends on an intact ubiquitin system. However, a role for direct ubiquitination of the substrate, as well as identification of potential ubiquitination sites, has not been discerned (Yu and Kopito, 1999). Similarly, ubiquitin-mediated endocytosis and degradation of the growth hormone receptor also proceeds in the absence of any lysine residue (Govers *et al*., 1999). The inability to identify ubiquitin adducts of the two receptors lead to the hypothesis that ubiquitination of another, yet to be identified factor, plays a role in the endocytic process.

Discovery of additional substrates is essential in order to establish N-terminal ubiquitination as a novel targeting pathway, to analyze the structural motifs involved, to identify the conjugating enzymes, and in particular the ubiquitin ligase, E3, and to study the physiological significance of this new pathway.

Here we show that HPV-16 E7 is a novel substrate of the ubiquitin pathway that is targeted for degradation via N-terminal ubiquitination.

Results

Degradation of E7 in a cell free reconstituted system requires ATP, formation of a polyubiquitin chain and the ubiquitin-carrier protein E2-F1

To study the mechanisms that underlie the degradation of E7, we reconstituted a cell free proteolytic system. As can be seen in Figure 1a, degradation of the protein requires three components, ATP, ubiquitin, and the ubiquitin carrier protein (E2) E2-F1 (E2-F1 is the rabbit homolog of the human UbcH7). Omission of any one of these components from the reaction mixture, abolished degradation. To further study the mechanism of ubiquitin action, we investigated whether formation of a polyubiquitin chain is required to promote degradation. To that end, we used the methylated derivative of ubiquitin that can modify the target protein only once and serves as a chain terminator (Hershko and Heller, 1985). As can be seen in Figure 1b, MeUb strongly inhibited degradation of E7 in the cell free system. This inhibition can be alleviated by the addition of excess free WT ubiquitin. To demonstrate directly generation of a substrate anchored polyubiquitin chain, we incubated labeled E7 in crude HeLa cell extract in the absence or presence of ATPγS. This nucleotide can support the activity of the ubiquitin activating enzyme E1 (in which the α-β bond is utilized), but not the activity of the 26S proteasome that requires cleavage of the β-γ bond (Johnston and Cohen, 1991). As can be seen in the experiment depicted in Figure 1c, incubation of labeled E7 in the presence of ATPγS generates a polyubiquitin chain that is anchored to the substrate.

Degradation of E7 in cells is mediated by the proteasome

To study the mechanism(s) that underlie degradation of E7 *in vivo*, we followed the stability of the protein in cells in the absence or presence of the specific proteasome inhibitor lactacystin. As can be seen in Figure 2, E7 is a short lived protein. Measurements in different experiments have demonstrated that the half life of the protein is $\sim 30-40$ min (see also Figure 7). Here, after 1 h, more than 70% of the protein is degraded. Addition of lactacystin inhibited degradation completely.

Degradation of a lysine-less E7 in a cell free reconstituted system requires ATP, formation of a polyubiquitin chain and the ubiquitin-carrier protein E2-F1

We have previously shown that degradation of the transcriptional activator MyoD requires attachment of the first ubiquitin moiety to the N-terminal free amino acid residue and not to any internal Lys residue of the protein (Breitschopf *et al*., 1998). Similarly, ubiquitin-mediated degradation of the Latent Membrane Protein 1 (LMP1) of the Epstein-Barr Virus is not dependent

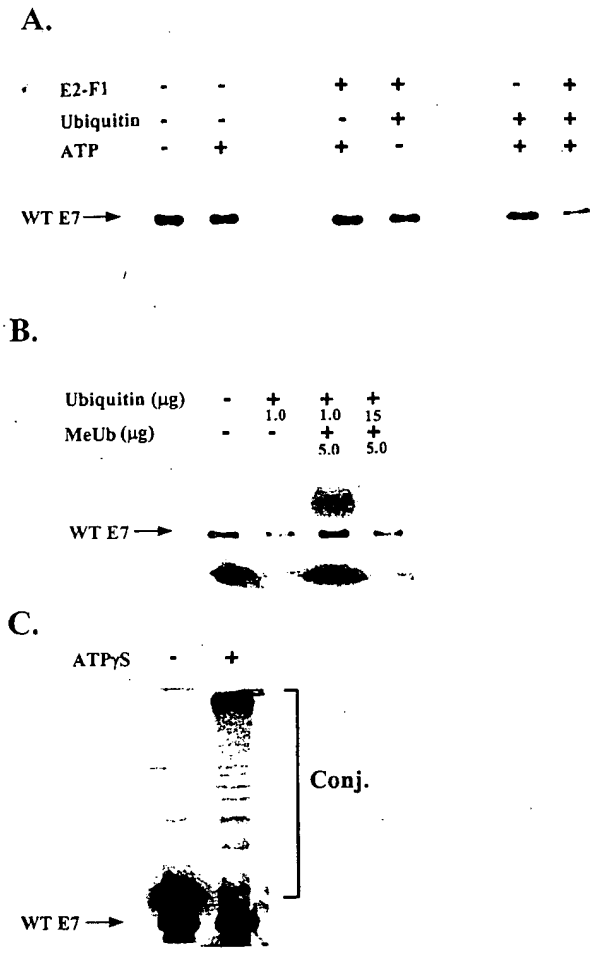


Figure 1 E7 is conjugated to ubiquitin and degraded *in vitro* in an ATP-, E2-F1-, and ubiquitin-dependent manner. (a) Degradation of E7 requires ATP, ubiquitin and E2-F1. Degradation of *in vitro* translated and ^{35}S-methionine labeled E7 was monitored in a cell free reconstituted system that contained reticulocyte Fraction II as described under Materials and methods. Ubiquitin, ATP, and E2-F1 were added as indicated. To avoid contamination of the labeled substrate with E2-F1, it was fractionated over DEAE prior to its addition to the reaction mixture as described under Materials and methods. (b) Degradation of E7 requires ubiquitin and formation of a polyubiquitin chain. Degradation of *in vitro* translated and labeled E7 was monitored in a cell free reconstituted system that contained HeLa cell Fraction II and ATP as described under Materials and methods. Ubiquitin, and MeUb were added as indicated. (c) Conjugation of ubiquitin to E7. *In vitro* translated and labeled E7 was incubated in complete HeLa cell extract in the absence or presence of ATPγS as indicated and as described under Materials and methods

on the single internal Lys residue of the molecule, and also requires initial fusion of ubiquitin to the N-terminal residue (Aviel *et al.*, 2000). To study whether a similar mechanism is also involved in the degradation of E7, we replaced the two Lys residues in positions 60 and 97 with Arg. As can be seen in the experiment depicted in Figure 3, similar to the WT protein, degradation of the lysine-less E7 in a cell free reconstituted system also requires ubiquitin, E2-F1, and ATP (Figure 3a). Degradation requires formation of a polyubiquitin chain (Figure 3b,c). As noted for MyoD and LMP1, the amount of ubiquitin adducts formed is lower in the case of the lysine-less mutant (compare c in Figures 1 and 3), suggesting that in the WT protein, internal Lys residues can also play a role, though not an essential one, in the proteolytic process.

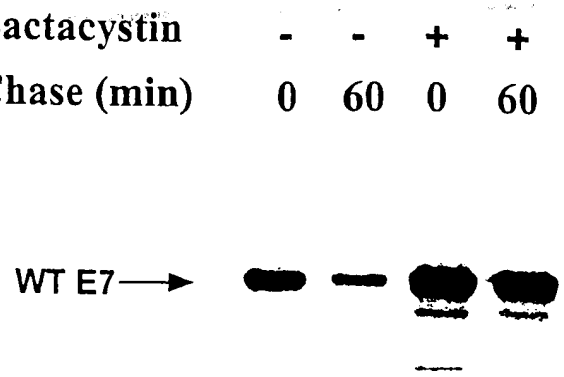


Figure 2 Degradation of E7 in cells is sensitive to proteasome inhibition. Cos 7 cells were transiently transfected with cDNA coding for E7. Degradation of the protein was monitored in a pulse-chase labeling and immunoprecipitation experiment in the absence or presence of the cell permeable proteasome inhibitor *clasto*-lactacystin β-lactone as described under Materials and methods

Degradation of lysine-less E7 in cells depends on an active proteasome

Similar to the WT protein, degradation of the lysine-less mutant in cells is also mediated by the proteasome (Figure 4).

Blocking of the N- but not of the C-terminus of E7 inhibits both conjugation and degradation of E7 both in vitro *and* in vivo

To study the involvement of the N-terminal domain in targeting the protein for N-terminal residue ubiquitination, we fused a 6 × Myc tag to the N-terminal and the C-terminal residues of E7. We predicted that if the N-terminal domain is involved in specific recognition of the protein, and ubiquitin will be fused only to a certain amino acid sequence, moving of this sequence downstream from the N-terminal domain will inhibit both conjugation and degradation. Indeed, as can be seen in Figure 5, conjugation of an N-terminally Myc-tagged WT E7 (that contains the two internal Lys residues) is strongly inhibited compared to a similar protein that contains a Myc tag fused to its C-terminal residue (Figure 5a). Similarly, while the degradation of a C-terminally Myc-tagged E7 in a cell free system proceeds efficiently, the degradation of an N-terminal Myc-tagged WT protein is strongly inhibited (Figure 5b). Not surprisingly, the degradation of a similar Lysine-less mutant is also inhibited (Figure 5b). We noted that the C-terminally tagged mutant migrates slower than its N-terminally tagged counterparts; the reason for this peculiar behavior is not known. Similarly, in cells, N-terminally tagged WT and lysine-less E7s are stable. unlike the C-terminally tagged WT protein (Figure 6).

The N-terminal domain of E7 contains the ubiquitination signal

To study directly the role of the N-terminal domain of E7 as a ubiquitination signal, we deleted the first 11 or 7 amino acids of the protein. As can be seen in Figure 7a, deletion of the first 11 residues stabilizes the

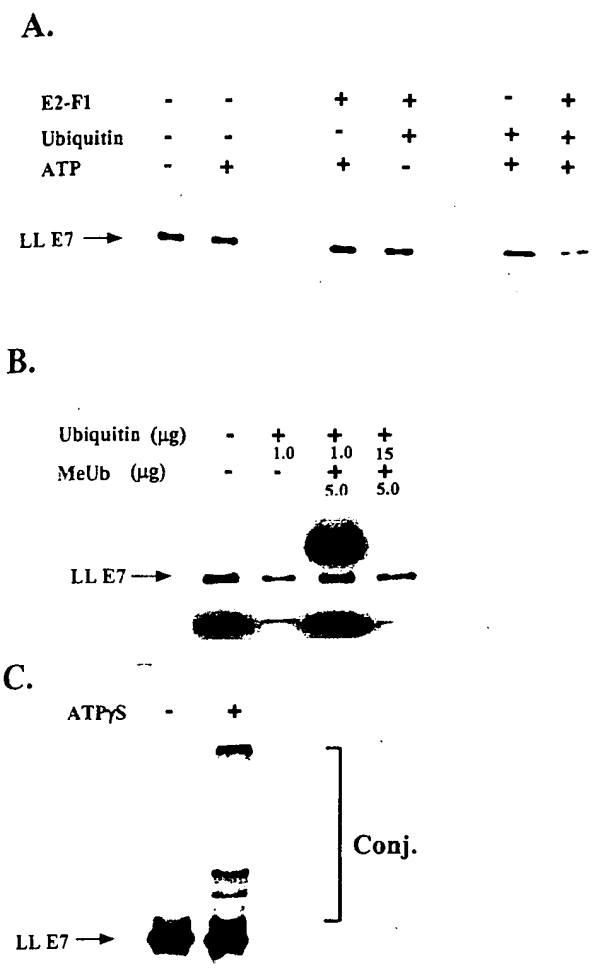


Figure 3 Lysine-less E7 is conjugated to ubiquitin and degraded *in vitro* in an ATP-, E2-F1-, and ubiquitin-dependent manner. (a) Degradation of lysine-less E7 requires ATP, ubiquitin and E2-F1. Degradation of *in vitro* translated and ^{35}S-methionine labeled lysine-less E7 was monitored in a cell free reconstituted system that contained reticulocyte Fraction II as described under Materials and methods. Ubiquitin, ATP, and E2-F1 were added as indicated. To avoid contamination of the labeled substrate with E2-F1, it was fractionated over DEAE prior to its addition to the reaction mixture as described under Materials and methods. (b) Degradation of lysine-less E7 requires ubiquitin and formation of a polyubiquitin chain. Degradation of *in vitro* translated and labeled lysine-less E7 was monitored in a cell free reconstituted system that contained HeLa cell Fraction II and ATP as described under Materials and methods. Ubiquitin, and MeUb were added as indicated. (c) Conjugation of ubiquitin to lysine-less E7. *In vitro* translated and labeled lysine-less E7 was incubated in complete HeLa cell extract in the absence or presence of ATPγS as indicated and as described under Materials and methods

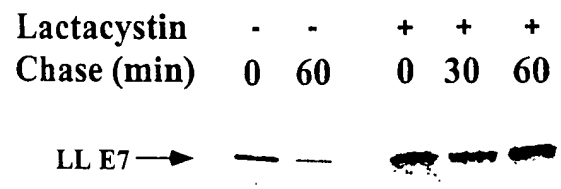


Figure 4 Degradation of lysine-less E7 in cells is sensitive to proteasome inhibition. Cos 7 cells were transiently transfected with cDNA coding for lysine-less E7. Degradation of the protein was monitored in a pulse-chase labeling and immunoprecipitation experiment in the absence or presence of the cell permeable proteasome inhibitor *clasto*-lactacystin β-lactone as described under Materials and methods

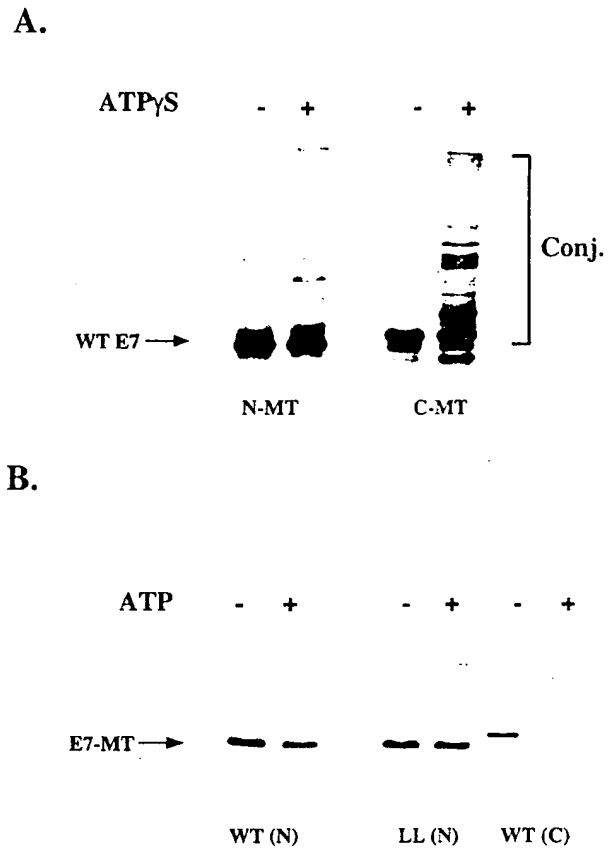


Figure 5 Conjugation (a) and Degradation (b) of N- and C-terminally Myc-tagged WT and lysine-less E7 *in vitro*. (a) ATPγS-dependent conjugation of N- and C-terminally Myc-tagged WT E7. N-terminally (N-MT) and C-terminally (C-MT) Myc-tagged *in vitro* translated and ^{35}S-methionine-labeled WT-E7 were subjected to ATPγS-driven conjugation in a complete HeLa cell extract that contains also ubiquitin. Reactions were carried out and conjugates resolved via SDS–PAGE as described under Materials and methods. (b) Degradation of N- and C-terminally Myc-tagged WT, and N-terminally Myc-tagged lysine-less E7 in a cell free system. WT [WT(N)] and lysine-less [LL (N)] N-terminally Myc-tagged and WT [WT (C)] C-terminally Myc-tagged E7 were subjected to degradation in a cell free system containing complete HeLa cell extract, ubiquitin, and ATP. Reactions were carried out and proteins resolved via SDS–PAGE as described under Materials and methods

protein. In contrast, degradation of a mutant protein from which the first 7 amino acid residues were deleted, proceeded similarly to that of the WT protein (Figure 7b). The finding that the signal is constituted of a relatively long segment raises the hypothesis that it serves not only as an anchor for specific ubiquitination, but also as a binding site for the ubiquitin ligase E3.

Discussion

Ubiquitin-mediated degradation of multiple key regulatory proteins is involved in the regulation of many basic cellular processes. Here we show that the HPV E7 oncoprotein is a substrate for the ubiquitin system. It is degraded in an ATP-dependent manner in a process that requires also ubiquitin and the ubiquitin-carrier protein E2-F1/UbcH7 (Figure 1a). It is not clear whether this E2 is the only carrier protein involved, or whether other E2s, such as members of the UbcH5 family that are involved in the degradation of the bulk

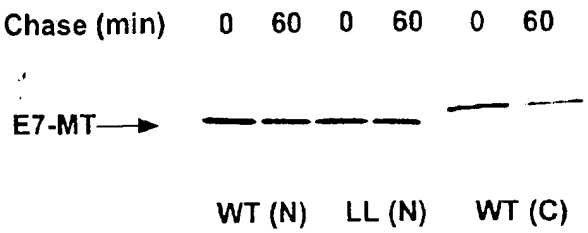


Figure 6 Degradation of N-terminally Myc-tagged WT and lysine-less and C-terminally Myc-tagged WT E7 in cells. Cos 7 cells were transiently transfected with cDNAs coding for WT [WT-MT(N)] and lysine-less [LL-MT(N)] N-terminally Myc-tagged and C-terminally [WT-MT(C)] Myc-tagged E7. Stability of the proteins was monitored in a pulse-chase labeling and immunoprecipitation experiment as described under Materials and methods

of cellular proteins, are also required. The identity of the ubiquitin-protein ligase E3 that binds the substrate specifically is still obscure. Degradation of E7 requires formation of a polyubiquitin chain: the chain terminator methylated ubiquitin inhibits degradation of E7, and inhibition can be relieved by the addition of excess free WT ubiquitin (Figure 1b). Indeed, we were able to demonstrate directly formation of polyubiquitinated E7 following incubation of the protein in the presence of ATPγS (Figure 1c). This nucleotide promotes formation of conjugates, but inhibits their degradation. In cells, E7 is extremely unstable and has a $t_{1/2}$ of $\sim$ 30–40 min. Incubation of cells in the presence of lactacystin, a specific proteasome inhibitor, inhibited degradation almost completely (Figure 2).

To further dissect the mechanisms that underlie the recognition and degradation of E7, it was important to study whether a specific Lys residue is essential for formation of the polyubiquitin chain, or whether the two Lys residues in positions 60 and 97 can equally serve as anchors to the polyubiquitin chain. We individually replaced each of the two Lys residues with Arg, however, we could not observe any effect on either ubiquitin-mediated conjugation and degradation *in vitro*, or on the stability of the protein *in vivo* (not shown). Therefore, we decided to replace these two residues. However, when incubated in a cell free reconstituted system, and similar to the behavior of the WT protein, degradation of the lysine-less E7 was still dependent on ATP, E2-F1, and ubiquitin (Figure 3). Also, in cells, degradation of the lysine-less protein was sensitive to inhibition of the proteasome (Figure 4). As no lysine residues were available for ubiquitination, we suspected that the first ubiquitin residue is attached to the free N-terminal NH_2 group, as is the case for MyoD (Breitschopf *et al.*, 1998) and EBV LMP1 (Aviel *et al.*, 2000). To study the possibility that E7 is also targeted via initial N-terminal ubiquitination, and with the assumption that the N-terminal domain of the protein determines the specificity for this process, we altered the N terminal domain of both WT and the lysine-less E7 by fusing it with a Myc-tag. For both forms, tagging resulted in major stabilization of the protein *in vitro* (Figure 5b) and *in vivo* (Figure 6). Concomitantly, we noted a marked decrease in conjugation of the tagged WT protein in a cell free system (Figure 5a). This decrease occurred despite the presence of the two internal Lys residues in the

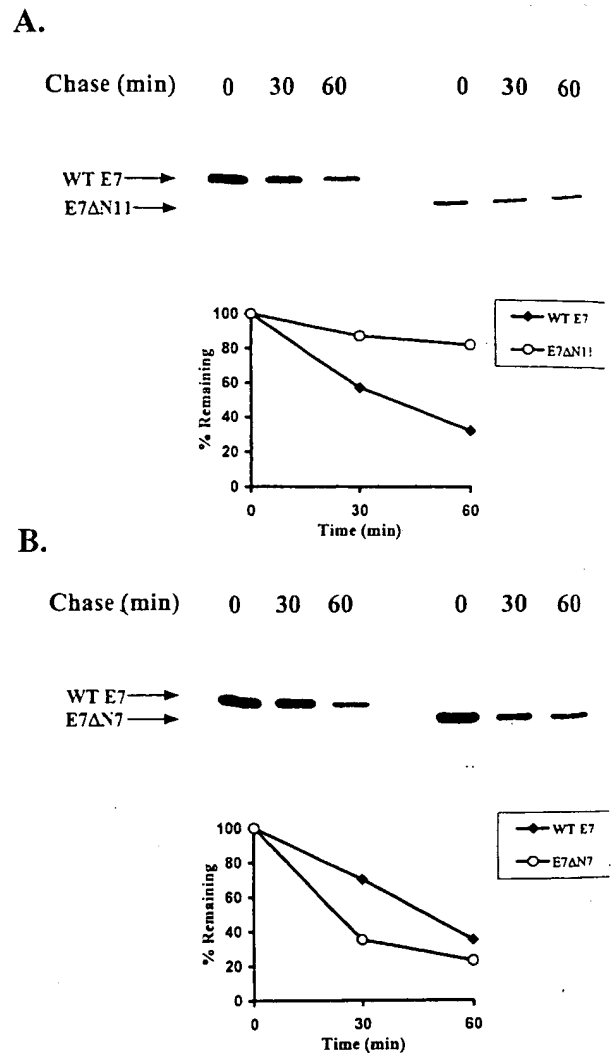


Figure 7 Degradation of 11 (a) and 7 (b) amino acid residues N-terminally deleted E7s in cells. (a) Degradation of 11 amino acid residues N-terminally deleted WT E7. Cos 7 cells were transiently transfected with a cDNA that codes for E7 lacking the first 11 amino acid residues (residues 2–12; ΔN11). Stability of the WT and the deletion mutant were monitored in a pulse-chase labeling and immunoprecipitation experiment as described under Materials and methods. Band intensities were quantified by analysis of the imaging data and plotted as relative percentage of the signal at time 0. (b) Degradation of 7 amino acid residues N-terminally deleted WT E7. Cos 7 cells were transiently transfected with a cDNA that codes for E7 lacking the first 7 amino acid residues (residues 2–8; ΔN7). Stability of the WT and the deletion mutant and quantitative analysis of the data were carried out as described under Materials and methods and above

protein. The effects of Myc tagging on both conjugation and degradation are even more striking if one takes into consideration the existence of additional six lysine residues in the tag. The finding that the Myc-tagged WT E7 can still be conjugated is probably due to the existence of the two internal lysine residue. Interestingly, while after 60 min the protein appears stable (Figure 6), it is still degraded in a much slower rate in a ubiquitin dependent mode (not shown). This finding demonstrates that the internal lysine residues (either of the protein and/or of the tag) can play a role in the process, though not a major one. In the native protein, they may serve as modulators of proteolysis. A similar observation was noted also for MyoD and

LMP1. The Myc tag can affect the stability of a protein that is targeted for degradation following ubiquitination of the N-terminal residue by blocking the access of ubiquitin, the ligase, or both, to a specific ubiquitination site and/or recognition motif at the N-terminal domain. To test this notion directly, we deleted the first 7 and 11 N-terminal amino acid residues. As can be seen in Figure 7, deletion of the first 11 residues stabilized the protein significantly. In contrast, deletion of the first seven residues was not sufficient to confer stability. Our initial results indicate that deletion of amino acid residues 8–11 results in stabilization of the protein, suggesting that they may play an important role in the recognition and targeting of the protein (unpublished data). Obviously, it will be important to study whether this sequence can serve as a 'universal' transferrable N-terminal destabilizing element. It should be noted, however, that we could not find sequence homology between the N-terminal regions of E7, MyoD, LMP1, and TCRα (it is not clear that this protein is degraded via N-terminal ubiquitination; Yu and Kopito 1999), but such functional motifs or epitopes may be generated following folding of the protein rather than at the primary sequence level. It is quite possible that the motifs are different and targeted by different ligases. Future identification of the E3 will be necessary to resolve the question of whether the N-terminal region serves also as a recognition site for the ligase. In this context it is worth mentioning the case of the Cdk inhibitor $p21^{Cip1}$, where the researchers suggested that, similar to ODC, the protein is targeted by the proteasome in a process that does not involve ubiquitination (Sheaff *et al.*, 2000); a lysine-less N-terminally Myc-tagged protein is still unstable and degraded in a proteasome-dependent manner.

It should be emphasized that N-terminal ubiquitination is different from the recognition via the N-end rule (Varshavsky, 1996) where the protein is recognized via the N-terminal residue, but conjugation occurs on internal lysines.

It is clear that for N-terminal ubiquitination to occur, a protein must have a free and exposed N-terminal residue. Thus, proteins acetylated at the N-terminal residue cannot be targeted via this pathway. While 80% of cellular proteins are modified at the N-terminal residue by acetylation (Jörnvall, 1975), 20% bear free and exposed N-termini. It is possible that these proteins are targeted via N-terminal ubiquitination. Indeed, in agreement with the set of 'rules' that determines, according to the three N-terminal residues, whether a protein will be acetylated, E7 that has MHG as its first three residues, should not be modified. However, since the set of 'rules' was determined for an extremely limited number of proteins, mostly yeast proteins (Polevoda *et al.*, 1999; Boissel *et al.*, 1988; Huang *et al.*, 1987), it is not clear whether it is possible to predict which proteins will be subjected to N-terminal ubiquitination. The hypothesis however can be tested experimentally. Furthermore, manipulation of proteins by altering their N-terminal domain and rendering them susceptible or resistant to acetylation, can also corroborate or rule out the notion that N-terminal ubiquitination is the pathway of 'choice' for free and exposed N-termini proteins.

Materials and methods

Materials

Materials for SDS–PAGE and Bradford reagent were from Bio-Rad. A mixture of L-^{35}S-labeled methionine and cysteine for metabolic labeling, [^{35}S]methionine for *in vitro* translation, as well as pre-stained MW markers and immobilized Protein G, were obtained from Amersham Pharmacia Biotech. Tissue culture sera and media were from Biological Industries (Bet Hae'emek, Israel) or from Sigma. Antibody against E7 was from Santa Cruz. *clasto*-lactacystin β-lactone was from Calbiochem. Ubiquitin, dithiothreitol (DTT), adenosine-5′-triphosphate (ATP), phosphocreatine, creatine phosphokinase, 2-deoxyglucose and [Tris(hydroxymethyl)aminomethane] (Tris buffer), were from Sigma. Hexokinase and Fugene™ 6 transfection reagent were from Roche Molecular Biochemicals. Wheat germ extract-based transcription-translation coupled kit (TNT®) was from Promega. Restriction and modifying enzymes were from New England Biolabs. Oligonucleotides were synthesized by Biotechnology General. (Rehovot, Israel). All other reagents were of high analytical grade.

Methods

Cell lines Cos-7 cells grown at 37°C in Dulbecco's Modified Eagle's Medium supplemented with 10% fetal calf serum (FCS). All transfections were carried out using the Fugene™ reagent, and cells were analysed after 36–48 h.

Plasmids and construction of mutant cDNAs WT and lysine-less mutant E7 cDNAs were subcloned into the *Eco*RI–*Xba*I site of the pCS2 and pCS2+MT vectors (Breitschopf *et al.*, 1998; Aviel *et al.*, 2000). These vectors were used for both *in vitro* translation (under the control of SP6 RNA polymerase) and expression in mammalian cells. Point mutations in E7 were generated by site-directed mutagenesis using the QuickChange™ kit (Stratagene). Deletion of the first 7 (ΔN7) or 11 (ΔN11) N-terminal amino acid residues of E7 was carried out using PCR and specific primers. PCR products were digested with *Eco*RI and *Xba*I and ligated into the pCS2 vector. In-frame insertion of 6×Myc tag in the N or C termini of E7 was carried out using the pCS+MT vector and the appropriate PCR primers. Sequences of all constructs were confirmed using an automatic sequencing system (ABI 310).

Preparation and fractionation of crude reticulocyte lysate Reticulocytes were induced in rabbits and lysates were prepared as described (Hershko *et al.*, 1983). The lysate was fractionated over DEAE cellulose onto unadsorbed material (Fraction I) and high salt eluate (Fraction II) as described (Hershko *et al.*, 1983). E2-F1 was prepared from Fraction I as described (Blumenfeld *et al.*, 1994). HeLa cell extract was prepared by hypotonic lysis as described previously (Orian *et al.*, 2000) and fractionated as described above.

Conjugation and degradation of E7 in a cell free reconstituted system The E7 cDNAs were translated in the presence of [^{35}S]Methionine using wheat germ coupled transcription-translation extract (TNT®, Promega) and SP6 RNA polymerase. When indicated, the crude lysate that contains the labeled substrate was fractionated over DEAE resin to Fraction I and II as described above. The labeled substrate was contained in Fraction II. Conjugation and degradation assays in a cell-free systems were performed as described elsewhere (Breitschopf *et al.*, 1998; Aviel *et al.*, 2000). Briefly, reaction mixture contained in a final volume of 12.5 μl: 50 μg whole HeLa cell lysate proteins, or 50 μg reticulocyte Fraction II and 1 μg E2-F1 as indicated, 5 μg ubiquitin,

and ~25 000 CPM of *in vitro* translated labeled E7. Reactions were performed in the presence of 0.5 mM ATP and an ATP-regenerating system (10 mM phosphocreatine and 5 μg phosphocreatine kinase), or ATPγS (5 mM) as indicated. For depletion of ATP, 0.5 μg hexokinase and 20 mM deoxyglucose were added. When indicated, the chain terminator methylated ubiquitin (MeUb; Hershko and Heller, 1985) was added at 5.0 μg. In these reactions ubiquitin was present at 1.0 μg. To overcome the inhibition of MeUb, 15 μg of ubiquitin were added. Conjugation assays contained in addition 0.5 μg of the isopeptidase inhibitor ubiquitin aldehyde (UbAl; Hershko and Rose 1987). Degradation reactions were carried out at 37°C for 2 h, whereas conjugation assays were incubated at 37°C for 1 h. Reactions were terminated by the addition of sample buffer and resolved by SDS-PAGE (15%). E7 was visualized by PhosphorImager (Fuji, Japan)

Stability of proteins in vivo Cellular stability of E7 proteins was monitored in a pulse-chase labeling and immunoprecipitation experiments as described (Breitschopf *et al.*, 1998). The proteasome inhibitor *clasto*-lactacystin β-lactone (10 μM) was added 20 min prior to the end of the labeling period (pulse) and was present throughout the experiment. Following labeling, cells were harvested (time 0: pulse) or were further incubated for the indicated periods of time (chase). Cells were lysed, and the labeled proteins were precipitated using anti-E7 antibody. Immune complexes were collected using immobilized protein G. Following SDS-PAGE (15%), proteins were visualized by a PhosphorImager.

Protein concentration Protein concentration was determined according to Bradford (1976) using BSA as a standard.

Acknowledgments

This research was supported by grants from the Israel Cancer Society, a TMR grant from the European Community, the Foundation for Promotion of Research at the Technion, and a research grant administered by the Vice President of the Technion for Research (to A Ciechanover), the US-Israel Binational Science Foundation (BSF; to A Ciechanover and AL Schwartz), the German-Israeli Cooperation Project (DIP) and the Israel Science Foundation founded by the Israeli Academy of Sciences and Humanities - Centers of Excellence Program (to A Ciechanover and M Oren), the German-Israeli Foundation for Scientific Research and Development (GIF; to A Ciechanover and M Scheffner). Purchasing of the ABI 310 autosequencer was supported partially by a grant from the Israel Science Foundation founded by the Israeli Academy of Sciences and Humanities.

References

Antinore MJ, Birrer MJ, Patel D, Nader L and McCance DJ. (1996). *EMBO J.*, **15**, 1950-1960.

Arbeit JM, Munger K, Howley PM and Hanahan D. (1993). *Am. J. Pathol.*, **142**, 1187-1197.

Arroyo M, Bagchi S and Raychaudhuri P. (1993). *Mol. Cell. Biol.*, **13**, 6537-6546.

Auewarakul P, Gissmann L and Cidarregui A. (1994). *Mol. Cell. Biol.*, **14**, 8250-8258.

Aviel S, Winberg G, Masucci M and Ciechanover A. (2000). *J. Biol. Chem.*, **275**, 23491-23499.

Blumenfeld N, Gonen H, Mayer A, Smith CE, Siegel NR, Schwartz AL and Ciechanover A. (1994). *J. Biol. Chem.*, **269**, 9574-9581.

Boissel J-P, Kasper TJ and Bunn FH. (1988). *J. Biol. Chem.*, **263**, 8443-8449.

Boyer SN, Wazer DE and Band V. (1996). *Cancer Res.*, **56**, 4620-4624.

Bradford MM. (1976). *Anal. Biochem.*, **72**, 248-254.

Breitschopf K, Bengal E, Ziv T, Admon A and Ciechanover A. (1998). *EMBO J.*, **17**, 5964-5973.

Chen LP, Ashe S, Singhal MC, Galloway DA, Hellstrom I and Hellstrom KE. (1993). *Proc. Natl. Acad. Sci. USA*, **90**, 6523-6527.

Crook T, Morgenstern JP, Crawford L and Banks L. (1989). *EMBO J.*, **8**, 513-519.

Davies R, Hicks R, Crook T, Morris J and Vousden K. (1993). *J. Virol.*, **67**, 2521-2528.

Funk J, Waga S, Harry J, Espling E, Stillman B and Galloway D. (1997). *Genes Dev.*, **11**, 2090-2100.

Govers R, ten-Broeke T, van-Kerkhof P, Schwartz AL and Strous GJ. (1999). *EMBO J.*, **18**, 28-36.

Greenhalgh DA, Wang XJ, Rothnagel JA, Eckhardt JN, Quintanilla MI, Barber JL, Bundman DS, Longley MA, Schlegel R and Roop DR. (1994). *Cell Growth Differ.*, **5**, 667-675.

Hawley-Nelson P, Vousden KH, Hubbert NL, Lowy DR and Schiller JT. (1989). *EMBO J.*, **8**, 3905-3910.

Hershko A, Heller H, Elias S and Ciechanover A. (1983). *J. Biol. Chem.*, **258**, 8206-8214.

Hershko H and Heller H. (1985). *Biochem. Biophys. Res. Commun.*, **128**, 1079-1086.

Hershko A and Rose IA. (1987). *Proc. Natl. Acad. Sci. USA*, **84**, 1829-1833.

Hu TH, Ferril SC, Snider AM and Barbosa MS. (1995). *Int. J. Oncol.*, **6**, 167-174.

Huang S, Elliott RC, Liu P-S, Koduri RK, Weickmann JL, Lee J-H, Blair LC, Ghosh-Dastidar RA, Bradshaw RA, Bryan KM, Einarson B, Kendall RL, Kolacz KH and Saito K. (1987). *Biochemistry*, **26**, 8242-8246.

Jörnvall H. (1975). *J. Theor. Biol.*, **55**, 1-12.

Johnston NL and Cohen RE. (1991). *Biochemistry*, **30**, 7514-7522.

Kanda T, Furuno A and Yoshiike KJ. (1988). *J. Virol.*, **62**, 610-613.

Kornitzer D and Ciechanover A. (2000). *J. Cell. Physiol.*, **182**, 1-11.

Munger K, Phelps WC, Bubb V, Howley PM and Schlegel R. (1989). *J. Virol.*, **63**, 4417-4421.

Orian A, Gonen H, Bercovich B, Fajerman I, Eytan E, Israël A, Mercurio F, Iwai K, Schwartz AL and Ciechanover A. (2000). *EMBO J.*, **19**, 2580-2591.

Pan H and Griep AE. (1994). *Genes Dev.*, **8**, 1285-1299.

Phelps WC, Yee CL, Munger K and Howley PM. (1988). *Cell*, **53**, 539-547.

Polevoda B, Norbeck J, Takakura H, Blomberg A and Sherman F. (1999). *EMBO J.*, **18**, 6155-6168.

Scheffner M, Werness BA, Huibregtse JM, Levine A and Howley PM. (1990). *Cell*, **63**, 1129-1136.

Selvey LA, Dunn LA, Tindle RW, Park DS and Frazer IH. (1994). *J. Gen. Virol.*, **75**, 1647-1653.

Sheaff RJ, Singer JD, Swanger J, Smitherman M, Roberts JM and Clurman BE. (2000). *Mol. Cell*, **5**, 403-410.

Varshavsky A. (1996). *Proc. Natl. Acad. Sci. USA*, **93**, 12142-12149.

Voges D, Zwickl P and Baumeister W. (1999). *Annu. Rev. Biochem.*, **68**, 1015-1068.

Yu H and Kopito RR. (1999). *J. Biol. Chem.*, **274**, 36852-36858.

Zerfass-Thome K, Zwerschke W, Mannhardt B, Tindle R, Botz J and Jansen-Durr P. (1996). *Oncogene*, **13**, 2323-2330.

Zwerschke W, Mazurek S, Massimi P, Banks L, Eigenbrodt E and Jansen-Durr P. (1999). *Proc. Natl. Acad. Sci. USA*, **96**, 1291-1296.